AD________________

Award Number: W81XWH-06-1-0056

TITLE: Greater Philadelphia Bioinformatics Alliance (GPBA) 3rd Annual Retreat 2005

PRINCIPAL INVESTIGATOR: David W. Russell, Ph.D.

CONTRACTING ORGANIZATION: Pennsylvania State University
University Park, PA 16802-7000

REPORT DATE: November 2005

TYPE OF REPORT: Final Proceedings

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-11-2005	Final Proceedings	14 Oct 05 – 14 Nov 05

4. TITLE AND SUBTITLE
Greater Philadelphia Bioinformatics Alliance (GPBA) 3rd Annual Retreat 2005

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-06-1-0056

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David W. Russell, Ph.D.

E-Mail: rzn@gv.psu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Pennsylvania State University
University Park, PA 16802-7000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains color plates: All DTIC reproductions will be in black and white.

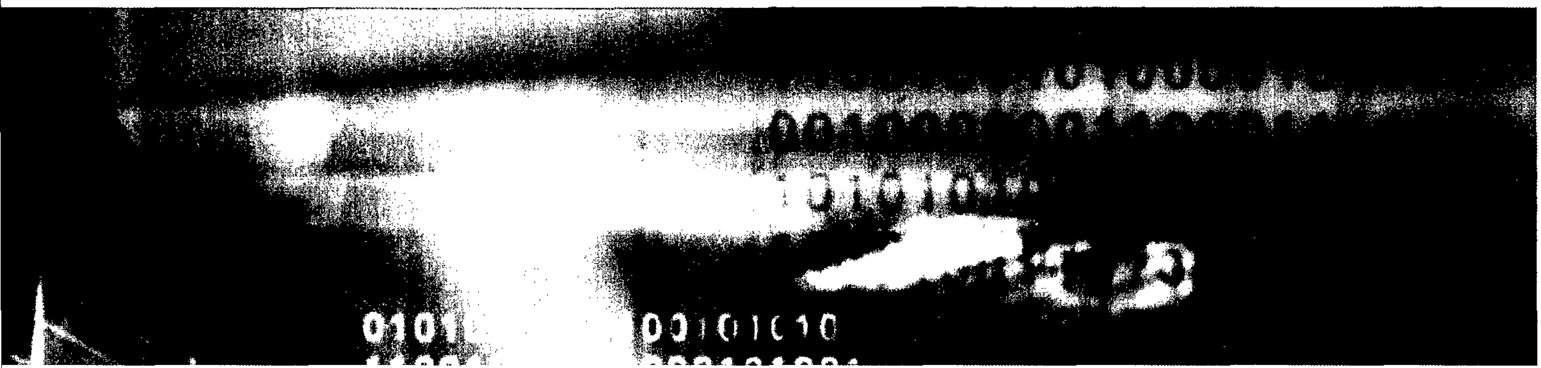
14. ABSTRACT

No abstract provided.

15. SUBJECT TERMS
No subject terms provided.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	114	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18



GPBA

Greater Philadelphia BioInformatics Alliance

3rd Annual Retreat
October 14, 2005
Penn State Great Valley

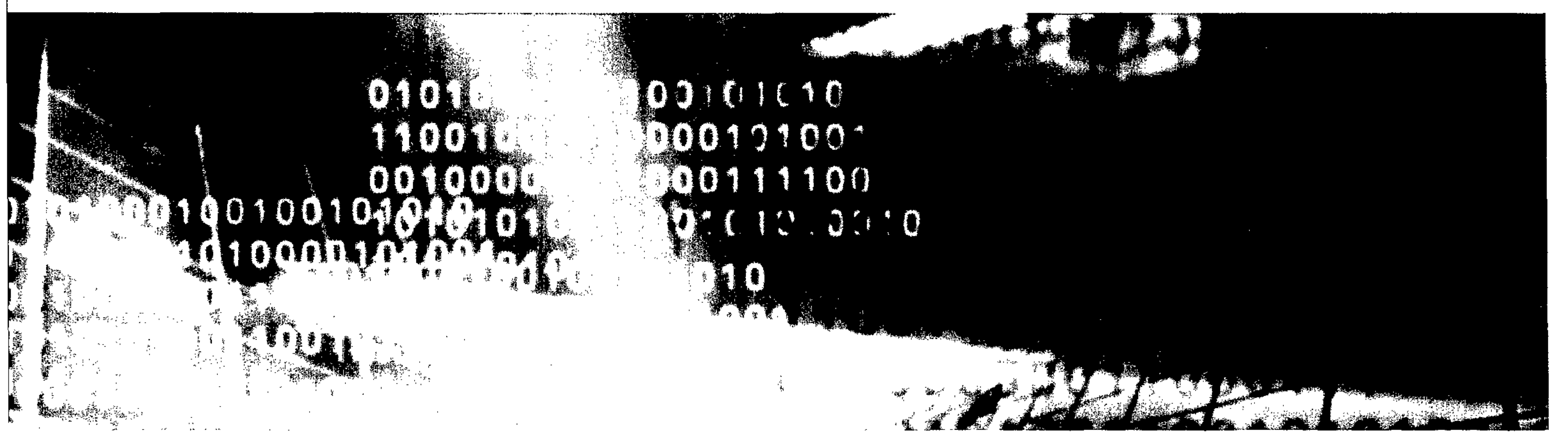
Proceedings

Retreat Sponsors

TATRC

Bio Advance

PROTECT · PROJECT · SUSTAIN



Greater Philadelphia Bioinformatics Alliance
3rd Annual Retreat
October 14, 2005

Welcome to the GPBA 3rd Annual Retreat. Please take a few minutes to provide feedback on your retreat experience and help us with ideas for future programs. Since your answers help us monitor and improve the programs we offer, all suggestions and comments are welcome.

Are you an:

- O Undergraduate student
- O Graduate student
- O Post-doc/Fellow
- O Faculty
- O Industry professional
- O Other ______________

From what discipline? ________________________________

Did you find the program topics useful and informative?

Session 1: Algorithms for "Omic" Analysis	Session 2: Biomedical Applications for Bioinformatics	Session 3: Industry Research	Panel Session: Academic/Industry Collaborations
O Very	O Very	O Very	O Very
O Somewhat	O Somewhat	O Somewhat	O Somewhat
O Not very	O Not very	O Not very	O Not very
O Not at all	O Not at all	O Not at all	O Not at all

Do you have comments about any particular topic or presentation? ______________________________________

____

If you could see anyone in the world speak about bioinformatics, computational biology, or related life sciences subjects, who would it be? If you could see anyone from Greater Philadelphia speak on these topics, who would it be?

____

What journals and publications do you read on a regular basis? Where do you go for science and industry news?

____

Do you have any suggestions to improve this event in the future?

____

Are you likely to attend the Annual Retreat next year? Circle one: Yes No

If you would like to be added to our mailing list, provide your email address below.

____

3rd Annual Greater Philadelphia Bioinformatics Alliance Retreat

Welcome to the Third Annual Greater Philadelphia Bioinformatics Alliance (GPBA) retreat. The GPBA is a regional consortium committed to providing significant value to the public, industry, and scientific community. We believe that bioinformatics and computational biology will play a crucial role in the future of bioscience and medicine and are an essential part of advancing life sciences in Greater Philadelphia.

The Greater Philadelphia region has the resources and potential to be a national leader in life sciences. The GPBA is one of many efforts to advance this cause by developing networks of innovation among academia and industry. Similar networks have historically supported and enabled the growth of such renowned regions as Silicon Valley.

In early 2002, a group of regional leaders came together with the goal of creating a multi-institutional "center of excellence" in bioinformatics. The effort was well received both by the regional universities and biomedical institutions. BioAdvance, the recently formed Biotechnology Greenhouse of Southeastern Pennsylvania, agreed to provide initial funding and development support to grow the Alliance. Information about the Greater Philadelphia Bioinformatics Alliance and its activities can be found at http://www.gpba-bio.com/.

The GPBA Annual Retreat represents one of the Alliance's programs developed to foster a connected research community. Designed to attract academic and industry researchers in bioinformatics and computational biology, the retreat highlights advanced research activity and offers opportunities for networking and exchange of information.

With that in mind, this year's retreat consists of technical talks on current research in bioinformatics, talks focused on research challenges faced by GPBA industry partners, a panel session entitled "Academic/Industry Collaborations in Bioinformatics and Computational Biology – Success Stories and Projects that Bombed: How?, Why?, Lessons Learned", a poster/demo session with expanded award categories, and a cocktail hour for networking and discussions.

Financial support for this retreat comes from TATRC and BioAdvance. We are grateful for their support, and for helping us provide the opportunity to showcase and strengthen the region's activity in bioinformatics and computational biology. We are also grateful to Pamela Vercellone-Smith for putting together the proceedings and making arrangements for the presentations.

We hope that you will take advantage of the proceedings, talks, posters, demos and lively discussions during the day to discover the wealth of bioinformatics research occurring in this region. We are glad that you have chosen to be part of this important event!

Thank you for attending and enjoy the meeting.

The retreat committee:
Susan B. Davidson (University of Pennsylvania)
Greg Gonye (Thomas Jefferson University)
Tammy Heesaaker (BioAdvance)
Guanghui Hu (GlaxoSmithKline)
David Russell (Penn State Great Valley)

GREATER PHILADELPHIA BIOINFORMATICS ALLIANCE 3rd ANNUAL RETREAT

OCTOBER 14, 2005

TABLE OF CONTENTS

I. PROGRAM and SCHEDULE OF EVENTS

8:30 -9:00	**Registration and Continental Breakfast**
8:55- 9:00	**Welcome to PSU Great Valley by Chancellor Disney**
9:00-9:05	**Opening and Logistics – Dr. Susan Davidson on behalf of GPBA**
9:05 -10:30	**Research Session 1: Algorithms for "Omics" Analysis (Session Chair: Susan Davidson)** Are there more signals lurking in genomic DNA? Isidore Rigoutsos (IBM) Biologically Motivated Pattern Recognition Techniques for Microarray Analysis. Michael Ochs (Fox Chase Cancer Center) An Efficient Computational Method to Identify a Protein Community from a Seed Protein. Daniel D. Wu and Xiaohua Hu (Drexel University)
10:30 -10:45	**Coffee Break**
10:45 -12:15	**Research Session 2: Biomedical Applications for Bioinformatics (Session Chair: Greg Gonye)** Designing and Mining (Pathogen) Genome Databases. David Roos (University of Pennsylvania) Transcriptional Regulatory Analysis of Retina Wound Healing Jerry Grunwald (Thomas Jefferson University) Computational Exploration of the Activated Pathways in Cancer Jan Feng (Temple University)
12:15 -1:30	**Lunch (with posters and demos set up)**

1:30 -3:00	**Industry Research - What Tough Bioinformatics/Life Sciences problems is industry tackling? (Session Chair: Guanghui Hu)**

Bioinformatics in Translational Science: Bridging the Gap between Pre-clinical and Clinical Discovery.

Anastasia M. Khoury Christianson, Ph.D., Director,
Discovery Medicine Informatics, AstraZeneca Pharmaceuticals

Target Identification through Expression Profiling and Pathway Analysis, A Case Study

Yuchen Bai, Ph.D., Senior Research Scientist,
Bioinformatics Client Services, Genomics, Wyeth Research

Integrative Biology: a mega-pixel view of disease and drug activity across species

Terence E Ryan, PhD, Director of Integrative Biology,
Discovery Research, GlaxoSmithKline

3:00 – 3:15	**Break**
3:15 -4:15	**Panel Discussion - Academic/Industry Collaborations: Success Stories and Projects that Bombed - How?, Why?, Lessons Learned (Moderator: Wade Rogers)**
4:15 -5:30	**Poster Session and Awards w/ Cocktails and Hors d'Oeuvres**

II. RESEARCH SESSION 1:

Algorithms for "Omics" Analysis

Session Chair: Susan Davidson, Ph.D.

9:05 – 10:30

RESEARCH SESSION 1:

Algorithms for "Omics" Analysis (Session Chair: Susan Davidson)

"Are there more signals lurking in genomic DNA?"

Isidore Rigoutsos

IBM

In this talk, I will describe recent work that we have been doing in my group and which revisits the question of what kinds of signals are present in genomic DNA. Our work seeks to explore the largely uncharted territory that lies beyond known exons and the regions immediately upstream of the known genes. In particular we have been focusing on two topics. First, we analyze computationally the phenomenon of post-transcriptional gene silencing in eukaryotic organisms. Our results so far suggest that the number of endogenously encoded microRNAs and the number of their targets may be substantially higher than is currently believed. Second, we examine the possibility that "logical" links exist between the non-coding and coding regions of genomic DNA. Here, our initial findings suggest the possibility that a wide range of biological processes is under the influence of a control layer which is currently unknown. Some results supporting the above two conjectures will be shown and discussed during the presentation.

Biologically Motivated Pattern Recognition Techniques for Microarray Analysis

Michael Ochs (Fox Chase Cancer Center)

Because of their ability to link genes that behave similarly across conditions, pattern recognition and clustering are widely used for data analysis in gene expression experiments. Numerous methods have been adapted or created to cluster genes or samples since initial microarray studies using hierarchical clustering. Developments have varied from adoption of standard clustering techniques, such as K-means clustering, to development of techniques targeted specifically to the expected issues in microarray data, such as QT-Clust and CAST. Generally these methods rely on placing each gene with a single cluster. Since biological systems reuse proteins for many different functions, genes actually belong more appropriately to multiple clusters. Some techniques, such as Biclustering, have been used to attempt to separate genes into multiple groups. However, these techniques generally rely on a gene being active in only one function in a single condition, which again is generally not true for living systems.

Two methods take a different approach, looking for models of the data that arise from linear combinations of behaviors. These look to solve the equation

$$D_{ij} = \sum_{i=1}^{k} A_{ik} P_{kj} + \varepsilon_{ij}$$

where D_{ij} are expression estimates for gene i in condition j, P_{kj} is the relative expression for pattern k in condition j, and A_{ik} is the relative strength of association of gene i with pattern k. The patterns $\boldsymbol{P}$ then provide links between conditions that ideally will be associated with biological processes, while the associations $\boldsymbol{A}$ show what genes are active in these processes.

We will present two approaches that solve this equation under constraints compatible with microarray data. Both systems rely on uncertainty levels estimated for each data point. The first, Bayesian Decomposition, relies on Bayesian statistics and Markov chain Monte Carlo simulation to estimate $\boldsymbol{A}$ and $\boldsymbol{P}$. The second, least squares nonnegative matrix factorization (LS-NMF), modifies NMF to handle cases where data points have different reliabilities. Results of application of these algorithms to microarray data will be presented in terms of ROC analysis.

An Efficient Computational Method to Identify a Protein Community from a Seed Protein

Daniel D. Wu and Xiaohua Hu
College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA
thu@cis.drexel.edu

Abstract--**Community structure is a topological property common to many networks. We present in this paper an efficient and accurate approach to detecting a community in a protein-protein interaction network from a given seed protein. Our experimental results show strong structural and functional relationships among member proteins within each of the communities identified by our approach, as verified by MIPS complex catalogue database and annotations.**

Index Terms-- **graph, community, protein-protein interactions**

I. Introduction

Proteins are important players in executing the genetic program. When carrying out a particular biological function, or serving as molecular building blocks for a particular cellular structure, proteins rarely act alone. Rather, biological complexity is encapsulated in the structure and dynamics of the combinatorial interactions among proteins (as well as other biological molecules) at different levels, ranging from biochemical pathways to ecological phenomena [1]. Therefore, one of the key challenges in this post genomic era is to understand these complex molecular interactions that confer the structure and dynamics of a living cell.

The development of high-throughput data collection techniques has generated tremendous amount of data about protein-protein interactions and molecular complexes, accompanied by intensive analyses of these large data sets in an attempt to understand and model the structure and dynamics of biological systems [2]. Modeling protein-protein interactions often takes the form of graphs or networks, where vertices represent proteins and edges represent the interactions between pairs of proteins. Research on such networks has revealed a number of distinctive topological properties, including the "small world effect", the power-law degree distribution, clustering (or network transitivity), and the community structure [3]. The "small world effect" refers to the fact of short average distance between vertices in a network. The degree (or connectivity) of a vertex in a network tells the number of other vertices that are connected to it. The degree distribution, $P(k)$, is the probability that a selected vertex has exactly degree of k. The degree distribution of most biological networks approximates a power law, $P(k) \sim k^{-\gamma}$, where γ is the degree exponent. Clustering refers to the phenomenon that if a vertex A is connected to vertex B, and B is connected to yet another vertex C, then there is a heightened probability that A also has a direct connection to C.

Community structure is another property common to many networks. Although there is no formal definition for the community structure in a network, it often loosely refers to the gathering of vertices into groups such that the connections within groups are denser than between groups [3]. The study of community structure in a network is not new. It is closely related to the graph partitioning in graph theory and computer science and the hierarchical clustering in sociology [4]. However, recent years have witnessed an intensive activity in this field partly due to the dramatic increase in the scale of networks being studied. Many algorithms for finding communities in networks have been proposed. They can be roughly classified into two categories, divisive and agglomerative. The divisive approach takes the route of recursive removal of vertices (or edges) until the network is separated into its components or communities, whereas the agglomerative approach starts with isolated individual vertices and joins together small communities. One important algorithm is proposed by Girvan and Newman (the GN algorithm) [3]. The GN algorithm is based on the concept of betweenness, a quantitative measure of the number of shortest paths passing through a given vertex (or edge). The vertices (or edges) with the highest betweenness are believed to play the most prominent role in connecting different parts of a network. The GN algorithm detects communities in a network by recursively removing these high betweenness vertices (or edges). It has produced good results and is well adopted by different authors in studies of various networks [4], but has a major disadvantage which is its computational cost. For sparse networks with n vertices, the GN algorithm is of n^3 time. Various alternative algorithms have been proposed [5-9], attempting to improve either the quality of the community structure or the computational efficiency of finding communities.

Because communities are believed to play a central role in the functional properties of complex networks [4], the ability to detect communities in networks could have practical applications. Studying the community structure of biological networks is of particular interest and challenging, given the enormous number of genes and proteins and the complex nature of interactions among them. In the context of biological networks, communities might represent structural or functional groupings, and can be synonymous with molecular modules,

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667)..

biochemical pathways, gene clusters, or protein complex. Being able to identify the community structure in a biological network hence could help us to understand better the structure and dynamics of biological systems. The GN algorithm has been applied to a number of metabolic networks from different organisms to detect communities that relate to functional units in the networks [10]. It has also been adapted to analyze a network of gene relationships as established by co-occurrence of gene names in published literature [11] and to detect communities of related genes.

In this paper, we address a slightly different question about the community structure in protein-protein interaction network, i.e. what is the community a given protein (or proteins) belongs to. Due to the complexity and modularity of biological networks, it is more feasible computationally to study a community containing one or a few proteins of interest. Hashimoto and colleagues [12] have used a similar approach to growing genetic regulatory networks from seed genes. Their work is based on probabilistic Boolean networks and sub-networks are constructed in the context of a directed graph using both the coefficient of determination and the Boolean function influence among genes. The similar approach is also taken by Flake and colleagues [13] to find highly topically related communities in the Web based on the self-organization of the network structure and on a maximum flow method.

Related works also include those that predict co-complex proteins. Jansen and colleagues [14] use a procedure integrating different data sources to predict the membership of protein complexes for individual genes based on two assumptions: first, the function of any protein complex depends on the functions of its subunits; and second, all subunits of a protein complex share certain common properties. Bader and Hogue [15] report a Molecular Complex Detection (MCODE) clustering algorithm to identify molecular complexes in a large protein interaction network. MCODE is based on local network density - a modified measure of the clustering coefficient. Bu and colleagues [16] use a spectral analysis method to identify the topological structures such as quasi-cliques and quasi-bipartites in a protein-protein interaction network. These topological structures are found to be biologically relevant functional groups. In our previous work, we developed a spectral-based clustering method using local density and vertex neighborhood to analyze the chromatin network [17-18]. Two recent works along this line of research are based on the concept of network modularity introduced by Hartwell and colleagues [19]. The works by [20] and [21] both used computational analyses to cluster the yeast protein-protein interaction network and discovered that molecular modules are densely connected with each other but sparsely connected with the rest of the network.

We present in this paper a new efficient and scalable computational method to discover the community that a given protein or proteins belong, based only on the topological properties, especially the community structure, of the protein-protein interaction network.

II. THE ALGORITHM

A. Notation

We intuitively model the protein-protein interaction network as an undirected graph, where vertices represent proteins and edges represent interactions between pairs of proteins.

An undirected graph, $G = (V, E)$, is comprised of two sets, vertices V and edges E. An edge e is defined as a pair of vertices (u, v) denoting the direct connection between vertices u and v. The graphs we use in this paper are undirected, unweighted, and simple – meaning no self-loops or parallel edges.

For a subgraph $G' \subset G$ and a vertex i belonging to G', we define the in-community degree for vertex i, $k_i^{in}(G')$, to be the number of edges connecting vertex i to other vertices belonging to G' and the out-community degree, $k_i^{out}(G')$, to be the number of edges connecting vertex i to other vertices that are in G but do not belong to G'.

In our algorithm, we adopt the quantitative definitions of community defined by Radicchi and colleagues [22], i.e. the subgraph G' is a community in a strong sense if $k_i^{in}(G') > k_i^{out}(G')$ for each vertex i in G' and in a weak sense if the sum of all degrees within G' is greater than the sum of all degrees from G' to the rest of the graph.

B. Algorithm

The algorithm, called *CommBuilder*, accepts the seed protein s, gets the neighbors of s, finds the core of the community to build, and expands the core to find the eventual community.

The two major components of *CommBuilder* are *FindCore* and *ExpandCore*. In fact, *FindCore* performs a naïve search for maximum clique from the neighborhood of the seed protein by recursively removing vertices with the lowest in-community degree until all vertices in the core set have the same in-community degree.

The algorithm performs a breadth first expansion in the core expanding step. It first builds a candidate set containing the core and all vertices adjacent to each vertex in the core. It then adds to the core a vertex that either meets the quantitative definition of community in a strong sense or the fraction of in-community degree over a relaxed affinity threshold f of the size of the core. The affinity threshold is 1 when the candidate vertex connects to each of vertices in the core set. This threshold provides flexibility when expanding the core, because it is too strict requiring every expanding vertex to be a strong sense community member.

The *FindCore* is a heuristic search for a maximum complete subgraph in the neighborhood N of seed s. Let K be the size of N, then the worst-case running time of *FindCore* is $O(K^2)$. The *ExpandCore* part costs in worst-case approximately $|V| + |E|$ + overhead. $|V|$ accounts for the expanding of the core, at most all vertices in V, minus what are already in the core, would be included. $|E|$ accounts for calculating the in- and out-degrees for the candidate vertices that are not in the core but in the neighborhood of the core. The overhead is caused by recalculating the in- and out-degrees of neighboring vertices every time the *FindCore* is recursively called. The number of these vertices is dependent on the size of the community we are building and the connectivity of the community to the rest of the network, but not the overall size of the network. For biological networks, the graphs we deal with are mostly sparse and small world, therefore, the running time of our algorithm will be close to linear.

III. EXPERIMENT RESULTS

To test our algorithm, we downloaded a dataset of interactions for *Saccharomyces cerevisae* from the General Repository for Interaction Datasets (GRID) [23]. The GRID database contains all published large-scale interaction datasets as well as available curated interactions such as those deposited in BIND [24] and MIPS [25]. The yeast dataset we downloaded has 4,907 proteins and 17,598 interactions.

We applied our algorithm against the network built from the downloaded dataset. The average running time for finding a community of 50 members is about 20 ms.

Algorithm 1 CommBuilder(G, s, f)

$G(V, E)$ is the input graph with vertex set V and edge set E.
s is the seed vertex, f is the affinity threshold.
$N \leftarrow$ {Adjacency list of s } $\cup$ {s}
$C \leftarrow$ FindCore(N)
$C' \leftarrow$ ExpandCore(C, f)
return C'

FindCore(N)
  for each $v \in N$
    calculate $k_v^{in}(N)$
  end for
  $Kmin \leftarrow \min\{k_v^{in}(N), v \in N\}$
  $Kmax \leftarrow \max\{k_v^{in}(N), v \in N\}$
  if $Kmin = Kmax$ **then return** N
  else return FindCore($N - \{v\}$, $k_v^{in}(N) = Kmin$)

ExpandCore(C, f)
  $D \leftarrow \bigcup_{(v,w)\in E, v\in C, w\notin C} \{v, w\}$
  $C' \leftarrow C$
  for each $t \in D$ **and** $t \notin C$
    calculate $k_t^{in}(D)$
    calculate $k_t^{out}(D)$
    if $k_t^{in}(D) > k_t^{out}(D)$ **or** $k_t^{in}(D)/|D| > f$ **then**
      $C' \leftarrow C' \cup \{t\}$
  end for
  if $C' = C$ **then return** C
  else return ExpandCore(C', f)

Because there is no alternative approach to our method, we decide to compare the performance of our algorithm to the work on predicting protein complex membership by Asthana and colleagues [26]. Asthana and colleagues reported results of queries with four complexes using probabilistic network reliability (we will refer their work as PNR method in the following discussion). Four communities are identified by CommBuilder using one protein as seed from each of the query complexes used by the PNR method. The seed protein is selected randomly from the "core" protein set. The figures for visualizing the identified communities are created using Pajek [27]. The community figures are extracted from the network we build using the above mentioned data set with out-of-community connections omitted. The proteins in each community are annotated with a brief description obtained from the MIPS complex catalogue database. As a comparison, we use Complexpander, an implementation of the PNR method [26] and available at http://llama.med.harvard.edu/Software.html, to predict co-complex using the core protein set that contains the same seed protein used by CommBuilder. For all our queries when using Complexpander, we select the option to use the MIPS complex catalogue database. We record the ranking of the members in our identified communities that also appear in the co-complex candidate list predicted by Comlexpander.

The first community, shown in Figure 1, is identified using TAF6 as seed. TAF6 is a component of the SAGA complex which is a multifunctional co-activator that regulates transcription by RNA polymerase II [28]. The SAGA complex is listed in MIPS complex catalogue as a known cellular complex consisting of 16 proteins. As shown in Table 1, the community identified by our algorithm contains 39 members, including 14 of the 16 SAGA complex proteins listed in MIPS (indicated by an asterisk in the *Alias* column). The community also contains 14 of 21 proteins listed in MIPS as Kornberg's mediator (SRB) complex. The rest of the proteins in the community are either TATA-binding proteins or transcription factor IID (TFIID) subunits or SRB related. TFIID is a complex involved in initiation of RNA polymerase II transcription. SAGA and TFIID are structurally and functionally correlated, make overlapping contributions to the expression of RNA polymerase II transcribed genes [28]. SRB complex is a mediator that conveys regulatory signals from DNA-binding transcription factors to RNA polymerase II [29]. In addition, 27 of the top 50 potential co-complex proteins (9 of the top 10), not including the seed proteins, predicted by Complexpander are in the identified community.

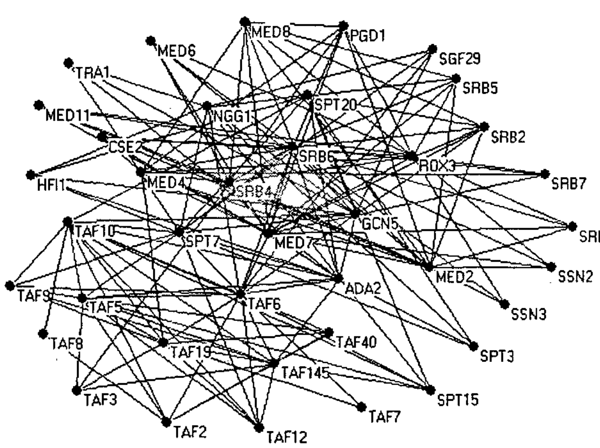


Fig. 1. The SAGA/SRB community.

TABLE I

THE SAGA/SRB COMMUNITY: PROTEINS THAT BELONG TO SAGA COMPLEX LISTED IN MIPS COMPLEX CATALOGUE DATABASE ARE INDICATED BY (*) AND THOSE BELONGING TO SRB COMPLEX ARE INDICATED BY (†). RANKING IS DONE BY RUNNING COMPLEXPANDER USING THE SEED (TAF6) AS THE CORE PROTEIN SET.

Protein	Alias	Description	Rank
YDR448w	ADA2*	general transcriptional adaptor or co-activator	1
YNR010w	CSE2†	subunit of RNA polymerase II mediator complex	
YGR252w	GCN5*	histone acetyltransferase	2
YPL254w	HFI1*	transcriptional coactivator	3
YMR112c	MED11†	mediator complex subunit	
YDL005c	MED2†	transcriptional regulation mediator	20
YOR174w	MED4†	transcription regulation mediator	23
YHR058c	MED6†	RNA polymerase II transcriptional regulation mediator	
YOL135c	MED7†	member of RNA Polymerase II transcriptional regulation mediator complex	21
YBR193c	MED8†	transcriptional regulation mediator	24
YDR176w	NGG1*	general transcriptional adaptor or co-activator	10
YGL025c	PGD1†	mediator complex subunit	37
YBL093c	ROX3†	transcription factor	
YCL010c	SGF29*	SAGA associated factor	43
YER148w	SPT15	the TATA-binding protein TBP	15
YOL148c	SPT20*	member of the TBP class of SPT proteins that alter transcription site selection	4
YDR392w	SPT3*	general transcriptional adaptor or co-activator	13
YBR081c	SPT7*	involved in alteration of transcription start site selection	5
YHR041c	SRB2†	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	
YER022w	SRB4†	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	27
YGR104c	SRB5†	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	
YBR253w	SRB6†	DNA-directed RNA polymerase II suppressor protein	19
YDR308c	SRB7†	DNA-directed RNA polymerase II holoenzyme and kornberg^s mediator (SRB) subcomplex subunit	46
YCR081w	SRB8	DNA-directed RNA polymerase II holoenzyme and Srb10 CDK subcomplex subunit	
YDR443c	SSN2	DNA-directed RNA polymerase II holoenzyme and Srb10 CDK subcomplex subunit	
YPL042c	SSN3	cyclin-dependent CTD kinase	
YGR274c	TAF1	TFIID subunit (TBP-associated factor), 145 kD	14
YDR167w	TAF10*	TFIID and SAGA subunit	7
YML015c	TAF11	TFIID subunit (TBP-associated factor), 40KD	18
YDR145w	TAF12*	TFIID and SAGA subunit	8
YML098w	TAF13	TFIID subunit (TBP-associated factor), 19 kD	17
YCR042c	TAF2	component of TFIID complex	22
YPL011c	TAF3	component of the TBP-associated protein complex	50
YBR198c	TAF5*	TFIID and SAGA subunit	9
YGL112c	TAF6*	TFIID and SAGA subunit	
YMR227c	TAF7	TFIID subunit (TBP-associated factor), 67 kD	
YML114c	TAF8	TBP Associated Factor 65 KDa	
YMR236w	TAF9*	TFIID and SAGA subunit	11
YHR099w	TRA1*	component of the Ada-Spt transcriptional regulatory complex	12

The second community is discovered using NOT3 as seed (Figure 2). NOT3 is a known component protein of the CCR4-NOT complex which is a global regulator of gene expression and involved in such functions as transcription regulation and DNA damage responses. MIPS complex catalogue lists 5 proteins for NOT complex and 13 proteins (including the 5 NOT complex proteins) for CCR4 complex. The NOT community identified is composed of 40 members. All 5 NOT complex proteins listed in MIPS and 11 of the 13 CCR4 complex proteins are members of the community. POL1, POL2, PRI1, and PRI2 are members of the DNA polymerase alpha (I) – primase complex, as listed in MIPS. RVB1, PIL1, UBR1, and STI1 have been grouped together with CCR4, CDC39, CDC36, and POP2 by systematic analysis [30]. The community also contains 20 out of 26 proteins of a complex that probably is involved in transcription and DNA/chromatin structure maintenance [31].

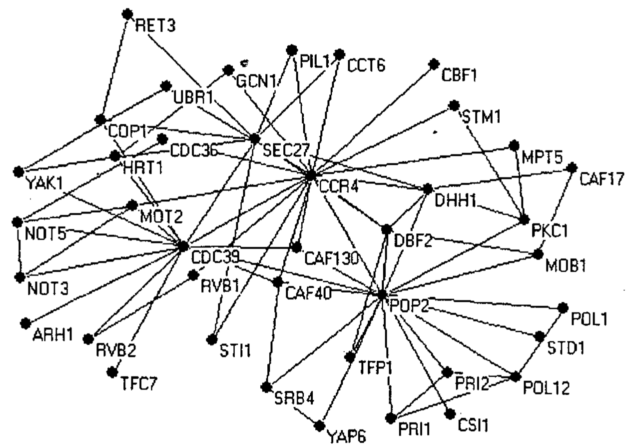


Fig. 2. The CCR4-NOT community.

TABLE II

THE CCR4-NOT COMMUNITY: PROTEINS BELONGING TO CCR4-NOT COMPLEX LISTED IN MIPS ARE INDICATED BY (*) AND PROTEINS CONSIDERED TO BE INVOLVED IN TRANSCRIPTION AND DNA/CHROMATIN STRUCTURE MAINTENANCE ARE INDICATE BY (†).

Protein	Alias	Description	Rank
YDR376w	ARH1	mitochondrial protein putative ferredoxin-NADP+ reductase	38
YGR134w	CAF130†	CCR4 Associated Factor 130 kDa	8
YJR122w	CAF17*	CCR4 associated factor	
YNL288w	CAF40†	CCR4 Associated Factor 40 kDa	9
YJR060w	CBF1	centromere binding factor 1	
YAL021c	CCR4*†	transcriptional regulator	3
YDR188w	CCT6†	component of chaperonin-containing T-complex (zeta subunit)	30
YDL165w	CDC36*†	transcription factor	40
YCR093w	CDC39*†	nuclear protein	1
YDL145c	COP1†	coatomer complex alpha chain of secretory pathway vesicles	11
YMR025w	CSI1	Subunit of the Cop9 signalosome, involved in adaptation to pheromone signaling	46
YGR092w	DBF2*	ser/thr protein kinase related to Dbf20p	6
YDL160c	DHH1*	DExD/H-box helicase, stimulates mRNA decapping,	17
YGL195w	GCN1†	translational activator	26
YOL133w	HRT1	Skp1-Cullin-F-box ubiquitin protein ligase (SCF) subunit	
YIL106w	MOB1*	required for completion of mitosis and maintenance of ploidy	10
YER068w	MOT2*†	transcriptional repressor	2
YGL178w	MPT5	multicopy suppressor of POP2	
YIL038c	NOT3*†	general negative regulator of transcription, subunit 3	
YPR072w	NOT5*†	component of the NOT protein complex	5

YGR086c	PIL1	Long chain base-responsive inhibitor of protein kinases Phk1p and Phk2p, acts along with Lsp1p to down-regulate heat stress resistance	
YBL105c	PKC1	ser/thr protein kinase	
YNL102w	POL1†	DNA-directed DNA polymerase alpha, 180 KD subunit	32
YBL035c	POL12†	DNA-directed DNA polymerase alpha, 70 KD subunit	28
YNR052c	POP2*†	required for glucose derepression	4
YIR008c	PRI1†	DNA-directed DNA polymerase alpha 48kDa subunit (DNA primase)	34
YKL045w	PRI2†	DNA-directed DNA polymerase alpha , 58 KD subunit (DNA primase)	31
YPL010w	RET3	coatomer complex zeta chain	39
YDR190c	RVB1	RUVB-like protein	29
YPL235w	RVB2†	RUVB-like protein	21
YGL137w	SEC27†	coatomer complex beta^ chain (beta^-cop) of secretory pathway vesicles	7
YER022w	SRB4	DNA-directed RNA polymerase II holoenzyme and Kornberg^s mediator (SRB) subcomplex subunit	44
YOR047c	STD1	dosage-dependent modulator of glucose repression	
YOR027w	STI1	stress-induced protein	
YLR150w	STM1	specific affinity for guanine-rich quadruplex nucleic acids	
YOR110w	TFC7†	TFIIIC (transcription initiation factor) subunit, 55 kDa	25
YDL185w	TFP1†	encodes 3 region protein which is self-spliced into TFP1p and PI-SceI	27
YGR184c	UBR1	ubiquitin-protein ligase	
YJL141c	YAK1	ser/thr protein kinase	
YDR259c	YAP6	transcription factor, of a fungal-specific family of bzip proteins	

The third community is identified by using RFC2 as the seed (Figure 3). RFC2 is a component of the RFC (replication factor C) complex, the "clamp loader", which plays an essential role in DNA replication and DNA repair. The community identified by our algorithm has 17 members. All five proteins of RFC complex listed in MIPS complex catalogue database are members of this community, as shown in Table 3. All but one member in this community are in the functional category of DNA recombination and DNA repair or cell cycle checkpoints according to MIPS. This community also includes the top 8 ranked proteins predicted by Complexpander.

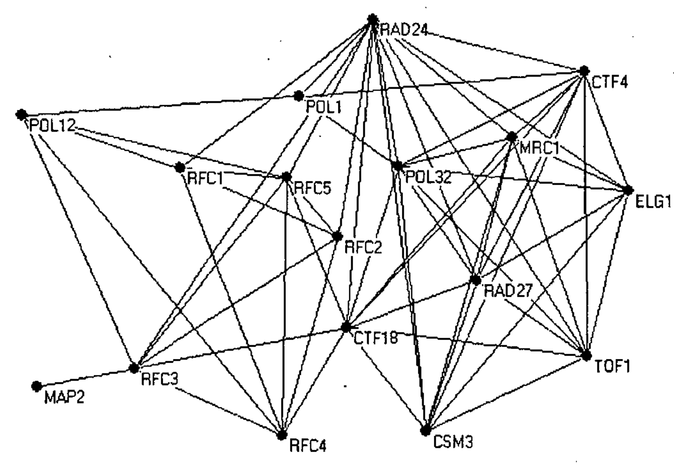


Fig. 3. The RFC community.

TABLE III

THE RFC COMMUNITY. PROTEINS BELONGING TO RFC COMPLEX LISTED IN MIPS ARE INDICATED BY (*) AND PROTEINS LISTED IN THE FUNCTIONAL CATEGORY OF DNA RECOMBINATION AND DNA REPAIR OR CELL CYCLE CHECKPOINTS BY MIPS ARE INDICATED BY (†).

Protein	Alias	Description	Rank
YMR048w	CSM3†	Protein required for accurate chromosome segregation during meiosis	
YMR078c	CTF18†	required for accurate chromosome transmission in mitosis and maintenance of normal telomere length	6
YPR135w	CTF4†	DNA-directed DNA polymerase alpha-binding protein	
YOR144c	ELG1†	Protein required for S phase progression and telomere homeostasis, forms an alternative replication factor C complex important for DNA replication and genome integrity	7
YBL091c	MAP2	methionine aminopeptidase, isoform 2	
YCL061c	MRC1†	Mediator of the Replication Checkpoint	
YNL102w	POL1†	DNA-directed DNA polymerase alpha, 180 KD subunit	19
YBL035c	POL12†	DNA-directed DNA polymerase alpha, 70 KD subunit	5
YJR043c	POL32†	polymerase-associated gene, third (55 kDa) subunit of DNA polymerase delta	
YER173w	RAD24†	cell cycle checkpoint protein	1
YKL113c	RAD27†	ssDNA endonuclease and 5^-3^exonuclease	
YOR217w	RFC1*†	DNA replication factor C, 95 KD subunit	8
YJR068w	RFC2*†	DNA replication factor C, 41 KD subunit	
YNL290w	RFC3*†	DNA replication factor C, 40 kDa subunit	2
YOL094c	RFC4*†	DNA replication factor C, 37 kDa subunit	4
YBR087w	RFC5*†	DNA replication factor C, 40 KD subunit	3
YNL273w	TOF1†	topoisomerase I interacting factor 1	

We use ARP3 as seed to identify the last community (Figure 5). ARP2/ARP3 complex acts as multi-functional organizer of actin filaments. The assembly and maintenance of many actin-based cellular structures likely depend on functioning ARP2/ARP3 complex [32]. The identified community contains all 7 proteins of the ARP2/ARP3 complex listed in MIPS (Table3). Not including the seed (ARP3), these proteins represent the top 6 ranked proteins predicted by Complexpander. As indicated in Table 4, there are 14 members belonging to the same functional category of budding, cell polarity, and filament formation, according to MIPS.

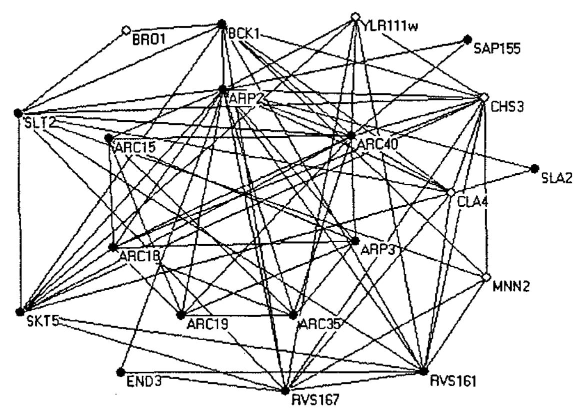


Fig. 4 The ARP2/ARP3 community.

TABLE IV

THE ARP2/ARP3 COMMUNITY. PROTEINS BELONGING TO ARP2/3 COMPLEX LISTED IN MIPS ARE INDICATED BY (*) AND PROTEINS LISTED IN THE FUNCTIONAL CATEGORY OF BUDDING, CELL POLARITY, AND FILAMENT FORMATION BY MIPS ARE INDICATED BY (†).

Protein	Alias	Description	Rank
YLR111w	YLR111w	hypothetical protein	
YIL062c	ARC15*†	subunit of the Arp2/3 complex	1
YLR370c	ARC18*	subunit of the Arp2/3 complex	4
YKL013c	ARC19*†	subunit of the Arp2/3 complex	3
YNR035c	ARC35*	subunit of the Arp2/3 complex	5
YBR234c	ARC40*†	Arp2/3 protein complex subunit, 40 kilodalton	6
YDL029w	ARP2*†	actin-like protein	2
YJR065c	ARP3*	actin related protein	
YJL095w	BCK1†	ser/thr protein kinase of the MEKK family	
YPL084w	BRO1	required for normal response to nutrient limitation	
YBR023c	CHS3†	chitin synthase III	
YNL298w	CLA4†	ser/thr protein kinase	
YNL084c	END3†	required for endocytosis and cytoskeletal organization	
YBR015c	MNN2	type II membrane protein	
YCR009c	RVS161†	protein involved in cell polarity development	
YDR388 w	RVS167†	reduced viability upon starvation protein	
YFR040w	SAP155†	Sit4p-associated protein	
YBL061c	SKT5†	protoplast regeneration and killer toxin resistance protein	
YNL243w	SLA2†	cytoskeleton assembly control protein	
YHR030c	SLT2†	ser/thr protein kinase of MAP kinase family	

IV. CONCLUSION AND FUTURE WORK

We present in this paper an efficient approach to growing a community from a given seed protein. It uses topological property of community structure of a network and takes advantage of local optimization in searching for the community comprising of the seed protein. Due to the complexity and modularity of biological networks, it is more desirable and computationally feasible to model and simulate a network of smaller size. Our approach builds a community of manageable size and scales well to large networks. Its usefulness is demonstrated by the experimental results that all the four communities identified reveal strong structural and functional relationships among member proteins. It provides a fast and accurate way to find a community comprising a protein or proteins with known functions or of interest. For those community members that are not known to be part of a protein complex or a functional category, their relationship to other community members may deserve further investigation which in turn may provide new insights.

Although we do not explicitly use our approach to the prediction of co-complexed proteins, the results of comparing with the PNR method developed by Asthana and colleagues [26] have shown that the communities identified by our approach do include the top ranked candidates of co-complexed proteins. Compared to the methods in predicting co-complexed proteins, our approach can discover a community rather than a single complex. In the context of this discussion, the notion of a community can be a complex, but it can also be a functional group consisting of several complexes, such as the SAGA/SRB community (Figure 1). This may not be always desirable. However, it does provide benefits of delineating the structure-function relationships beyond a single complex. In this spirit, one part of our future work is to further explore the relaxation threshold (*f*) aiming to identify either a more tightly connected community under a more strict expanding condition or a more loosely connected community under a relaxed condition so that we could study interactions of different strengths within a community.

Our approach does not consider the quality of data in our downloaded data set. By using the strong sense definition of community [22], we could to some degree reduce the noises. However, to improve the quality of an identified community, we have to take into account the quality of data and that is another part of our future work. One possible way is to use the probabilities assigned to individual protein pairs as used by [14], [22], [33], and [34].

REFERENCES

[1] Barabasi, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101–114.

[2] Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 14: 292–299.

[3] Girvan, M. and Newman, M.E.J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99: 7821-7826.

[4] Newman, M.E.J. (2003). The Structure and Function of Complex Networks. *SIAM Review* 45(2): 167-256

[5] Newman, M. E. J. (2004). Detecting community structure in networks. *Eur. Phys. J. B* 38: 321-330.

[6] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69: 026113.

[7] Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69: 066133.

[8] Donetti, L. and Munoz, M.A. (2004). Detecting Network Communities: a new systematic and efficient algorithm. *J. Stat. Mech.* P10012.

[9] White, S. and Smyth, P. (2005). A Spectral Clustering Approach to Finding Communities in Graphs. *SIAM International Conference on Data Mining* 2005, Newport Beach, CA, USA.

[10] Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19(4): 532-538.

[11] Wilkinson, D. and Huberman, B.A (2004). A Method for Finding Communities of Related Genes. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl 1): 5241-5248.

[12] Hashimoto,R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., and Dougherty, E.R. (2004). Growing genetic regulatory networks from seed genes. *Bioinformatics* 20(8): 1241–1247.

[13] Flake, G. W., Lawrence, S. R., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of Web communities, *IEEE Computer* 35: 66-71.

[14] Jansen, R., Lan, N., Qian, J., and Gerstein, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Functional Genomics* 2: 71–81.

[15] Bader, G.D. and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2.

[16] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G. and Chen, R. (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.* 31: 2443–2450.

[17] Hu, X. (2005). Mining and Analyzing Scale-free Protein-Protein Interaction Network, *International Journal of Bioinformatics Research and Application* 1(1): 81-101.

[18] Hu, X.; Yoo, I., Song, I.-Y., Song, M., Han, J. and Lechner, M. (2004). Extracting and Mining Protein-Protein Interaction Network from Biomedical Literature, in the *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (IEEE CIBCB 2004), Oct. 7-8, 2004, San Diego, USA (Best paper award).

[19] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999). From molecular to modular cell biology. *Nature* 402: C47–C52.

[20] Spirin, V. and Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.* 100: 12123–12128.

[21] Rives, A.W. and Galitski, T. (2003). Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U.S.A.* 100: 1128–1133.

[22] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U.S.A.* 101: 2658-2663.

[23] Breitkreutz, B.-J., Stark, C. and Tyers, M. (2003). The GRID: The General Repository for Interaction Datasets. *Genome Biology* 4: R23.

[24] Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31(1): 248-250.

[25] Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* 30: 31–34.

[26] Asthana, S., King, O.D., Gibbons, F.D., and Roth, F.P. (2004). Predicting Protein Complex Membership Using Probabilistic Network Reliability. *Genome Res.* 14: 1170-1175

[27] Batagelj, V. and Mrvar, A. (1998). Pajek: Program for large network analysis. *Connections* 21: 47–57.

[28] Wu, P.Y., Ruhlmann, C., Winston, F., and Schultz, P. (2004). Molecular architecture of the *S. cerevisiae* SAGA complex. *Mol. Cell* 15: 199–208.

[29] Guglielmi, B., van Berkum, N.L., Klapholz, B., Bijma, T., Boube, M., Boschiero, C., Bourbon, H.M., Holstege, F.C.P., and Werner, M. (2004). A high resolution protein interaction map of the yeast Mediator complex. *Nucleic Acids Res.* 32: 5379–5391.

[30] Ho, Y., et al (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415: 180 – 183.

[31] Gavin, A.-C., et al (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141 – 147.

[32] Machesky, L.M. and Gould, K.L. (1999). The Arp2/3 complex: a multifunctional actin organizer. *Curr. Opin. Cell Biol.* 11: 117 – 121.

[33] Bader, J.S. (2003). Greedily building protein networks with confidence. *Bioinformatics* 19(15): 1869-1874.

[34] Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *NATURE BIOTECH.* 22(1): 78-85

III. RESEARCH SESSION 2:

Biomedical Applications of Bioinformatics

Session Chair: Greg Gonye, Ph.D.

10:45 – 12:15

RESEARCH SESSION 2:

Biomedical Applications of Bioinformatics (Session Chair: Greg Gonye)

Designing and Mining (Pathogen) Genome Databases

David S. Roos et al.
Department of Biology, and Penn Genomics Institute
University of Pennsylvania, Philadelphia PA 19143-6018 USA

Genomic-scale projects yield vast datasets, from genome and EST sequences, to RNA and protein expression profiles, to interactome and metabolic pathway data, to polymorphisms identified at the population level, and comparative genomics data gleaned from cross-species analysis. Valuable though they may be, however, the emergence of such data -- at ever-increasing rates -- raises an important problem: how to effectively capture, maintain, update, annotate, integrate, and query these resources to advance biomedical research? Genome database development presents challenges for any organism, but certain consistent features apply to taxonomically diverse pathogen species. For example, in contrast to most studies on human metabolic diseases, highly abundant targets are often of greatest interest for drug/vaccine/diagnostic development. Taxonomically-related species permit revealing comparisons between pathogenic and non-pathogenic organisms, facilitating the development of broad-spectrum antibiotics. Correlations between pathogen and host genomes provide additional opportunities for productive exploration.

The Plasmodium genome database (PlasmoDB.org) provides access to information emerging from various genome sequencing and functional genomics projects for several parasite species, enabling malaria researchers to formulate their own queries. In 2004, PlasmoDB received >6M hits from >45K unique users in >100 countries. Data types available for browsing, downloading, analysis, and dynamic queries include genome and EST sequence for eight Plasmodium species, curated and automated analyses of gene/protein predictions, RNA and protein expression data, data on genetic mutability and population diversity, protein interactome data, ortholog/paralog identification, reagents, publication records, user comments, etc. Particular effort has been invested in developing and exploiting a strongly-typed relational schema for representing a wide range of complete and incomplete datasets. Integrating these diverse datasets, both within the database and in the form of user queries, enables various lines of evidence to be explored in order to identify alternative gene models, assess expression profiles, etc. Comparative and phylogenetic approaches include the automatic identification of orthologs and transitive assignment of probable annotation. Combining data from Plasmodium with the related human and veterinary pathogens Toxoplasma and Cryptosporidium yields an integrated apicomplexan parasite database (ApiDB), enabling cross-species comparisons. Comparison with human (and vector) genome(s) has expedited a variety of projects of biological and evolutionary interest, and highlights phylogenetically-restricted targets suitable for diagnostic, drug, and vaccine design.

In Vitro and in Silico Approaches to Understanding Proliferative Vitreoretinopathy

G.B. Grunwald, C.H. Pratt, R. Vadigepalli, and G.E. Gonye. Department of Pathology, Anatomy and Cell Biology, Jefferson Medical College, Thomas Jefferson University, 1020 Locust Street, Philadelphia, PA 19107.

The long-term objective of our research program is to enhance our understanding of the cellular and molecular mechanisms of retinal development and to apply this knowledge to the prevention and treatment of retinal diseases. The underlying philosophy is that nature is parsimonious, and that the principles of embryonic development, as exemplified by highly orchestrated cell-cell adhesive interactions and related signaling pathways, are applicable to events in adult tissues such as the normal wound healing response to injury. This also applies to the aberrant wound healing that can result in pathological changes in tissues such as the retina. The retinal pigment epithelium (RPE), whose intrinsic barrier properties as well as neural retinal interactions are both critical for the development and maintenance of normal visual function, provides a case in point. The RPE possesses a limited wound healing ability, and in diseases such as proliferative vitreoretinopathy (PVR) this process may be subverted resulting in the fibrotic retinal abnormalities associated with this disease. PVR is the leading cause of failure of surgical approaches to repair of retinal detachment. We envision PVR as an "epigenetic" disease that results from a maladaptive wound healing response resulting from a combination of factors including alteration in cadherin cell adhesion molecule expression and associated cell signaling and gene expression patterns. Our guiding hypothesis is that the aberrant wound healing that occurs among RPE cells during PVR results from an inappropriate epithelial-mesenchymal phenotypic switch that includes abnormal cadherin subtype expression. This phenotypic switch results from a subversion of the normal wound healing process and is stably reinforced via a cell signaling feedback loop generated by a specific combination of cadherin/cytokine receptor complex formation, resultant downstream cytoplasmic signaling, and ultimate changes in gene expression. In order to critically test this hypothesis, we are applying a combination of traditional cell and molecular biology as well as bioinformatics and computational biology approaches to identifying the cell signaling pathways that initiate these phenotypic changes, and the gene regulatory networks that are affected by, and in turn further modify and reinforce, the changes in cell state that accompany PVR. Using the transcriptional regulatory network analysis tool PAINT (Promoter Analysis and Interaction Network Toolset; www.dbi.tju.edu), we have begun to identify components of the genetic regulatory network that modulate RPE phenotype and cadherin subtype expression following wounding and vitreous-mediated phenotypic transformation. Through the combination of in vitro and in silico approaches, we have identified key steps at the cell surface, along intracellular signaling pathways, and at the transcriptional level, that will not only enhance our understanding of fundamental retinal cell biological processes, but will also provide potential therapeutic targets for the prevention and treatment of retinal diseases such as PVR. Supported by grants NIH RO1EY06658, NIHT32ES07282, and DARPA F30602-01-2-0578.

A Computational Exploration of the Activated Pathways in Cancer

Liting Wen, Wei Li and Jan Feng

Department of Chemistry, Temple University
feng@temple.edu

INTRODUCTION: The occurrence of DNA damage triggers cellular responses by the DNA repair system. Such responses include the recruitment of DNA repair proteins to the damaged site, cell cycle arrest, and apoptosis if the DNA repair machinery is not able to repair the damage[1]. How the cell controls the initiation of different cellular responses to the DNA damage signals is still the subject of many studies. One of the emerging themes is that the cellular response to DNA damage involves intimate coupling of the DNA repair systems and the cell-cycle regulation. Cells defective in cell-cycle checkpoints often fail to induce cell-cycle arrest when exposed to DNA-damaging agents[2]. Cells with deficient DNA repair systems also have elevated genetic mutation rates and often transform into neoplastic cells. Defects in the DNA repair system have been linked to a number of human diseases ranging from autosomal recessive diseases to sporadic DNA damage processing diseases and cancer.

In this report, we describe a bioinformatics approach to explore potentially activated signaling pathways associated with DNA damage response in cancer. We have developed RepairNET, an extended cellular network associated with the DNA damage response[3]. RepairNET served as an intrinsic framework from which the activated signaling pathways were identified. The principle of our method relied on the observation that genes with correlated expression profiles at mRNA level often implied coordinated functions of their gene products in various cellular processes[4]. We developed a pathway-exploration algorithm that systematically searched the expression data for genes corresponding to the proteins of RepairNET with significantly correlated expression profiles. These groups of proteins would form the activated signaling pathways associated with the DNA damage response. We performed an analysis on a previously published gene expression data of breast cancer. This dataset consisted of gene expression profiles for samples from patients who developed distant metastases within 5 years and patients with greater than 5 year survival. The results of our analysis revealed distinct activated pathways in each sampling category.

METHODS: RepairNET was derived from the Repair-FunMap, a database containing functional relationship of proteins associated with DNA repair. Repair-FunMap was a manually curated database. The database assembly was assisted by a literature data mining algorithm that searched the MEDLINE database for literature describing protein interactions related to DNA repair[3]. The initial entries to Repair-FunMap were derived from the content of a publicly available human DNA repair database that was expertly compiled[5]. These proteins were directly involved in the recognition and the repair of the damaged DNA. Two scores, the $S(Repair)$ and the $S(Interaction)$, were calculated for abstracts in the MEDLINE[3]. Abstracts having the property that both scores were above a pre-determined threshold were selected for manual curation. The manual curation step ensured the reliability of the database content. The names and the functional relationships of the proteins discussed in the abstracts were extracted and added to the initial human DNA repair database. The first iteration generated the first layer of known proteins interacting with the human DNA repair proteins. Repeated iterations of the described protocol resulted in an extended network associated with DNA damage response. This extended protein interaction network was a comprehensive collection of functional relationships of proteins associated with the DNA damage response. Currently, RepairNET contains 133 human DNA repair proteins and 1200 proteins that are involved in the DNA damage response with more than 2300 protein interactions. RepairNET is available online at http://guanyin.chem.temple.edu.

The pathway-exploration algorithm rested on two suppositions: (i) genes with causally connected expression profiles at the mRNA level often implied coordinated functions of their gene products in various cellular processes, and (ii) the expression profile of each gene in the activated pathways could be described, with reasonable accuracy, by regressing it on the expression profiles of those genes whose products were connected in RepairNET that caused it. Given RepairNET, the algorithm identified genes with significantly correlated expression profiles and, additionally, a Bayesian network that best characterized their mutual relationship. One solution to this optimal characterization involved searching RepairNET for groups of proteins whose corresponding gene expression profiles had significant multiple correlations subject to certain conditions.

RESULTS: We applied the pathway-exploration algorithm to analyze published microarray data of breast cancer. The breast cancer was of the sporadic type. The study employed two-channel oligonucleotide microarray containing 25,000 human genes, which measured gene expression profiles against a reference cRNA pool[6]. The microarray data contained 930 genes whose products overlapped with proteins in RepairNET. The differential log ratios of the gene expression profiles were used to calculate the correlation coefficients between various gene pairs.

Of the 78 samples in the study, 34 samples were from patients who developed distant metastases within 5 years (the metastasis group) and 44 samples were from patients who were diseases-free after 5 years (the disease-free group)[6]. We applied the pathway-exploration algorithm to identify the potentially activated signaling pathways in RepairNET for both sample groups. For the metastasis group, we found 75 pairs of proteins with gene expression profiles that were significantly correlated ($P_{\text{permuted}} < 0.005$ and $P_{\text{paired}} < 0.06$), of which 23 pairs formed 4 different pathways that contained at least 6 proteins (figure 1a). These four pathways were known to be associated with cell cycle checkpoints control, apoptosis, and DNA repair. For the microarray data of the disease-free group, we found 140 pairs of proteins in RepairNET with significantly correlated gene expression profiles. 4 signaling pathways were formed among 38 of these protein pairs (Figure 1b). The most prominent pathway, which contained 28 proteins, corresponded to the cellular processes associated with DNA repair activities. Some of the active pathways found in the metastasis group were also activated in this group of samples.

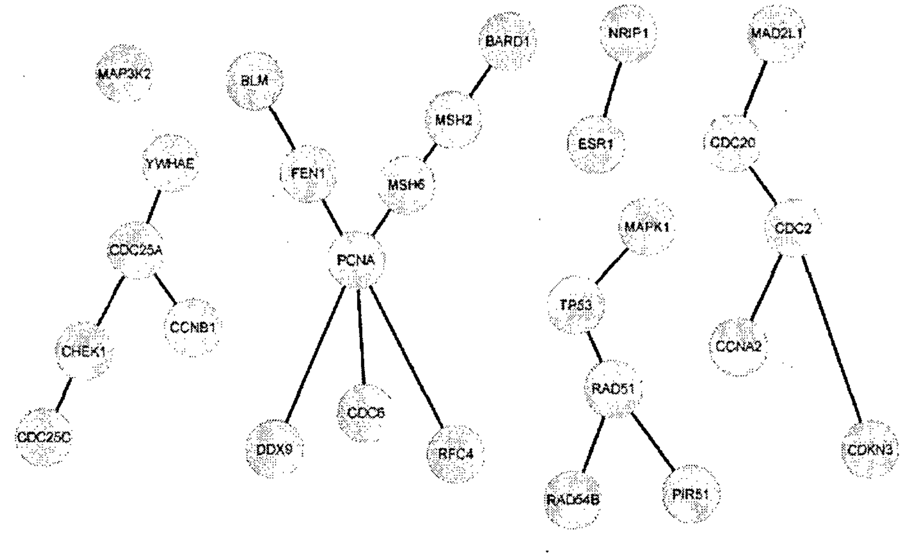


Figure 1a

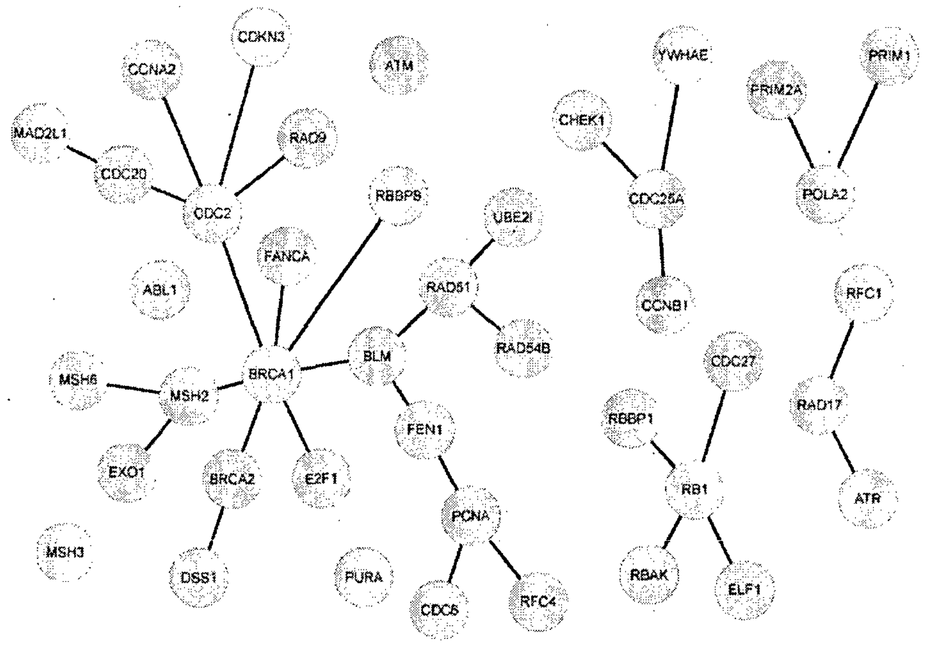


Figure 1b

Figure 1: Activated signaling pathways in breast cancer. The nodes are labeled with gene names using the Entrez Gene symbols. Dark lines connect genes with positively correlated expression profiles, and the light gray lines connect genes with negative correlations. (a) Four activated pathways derived from microarray data of the metastasis group; (b) four activated pathways derived from microarray data of the disease-free group.

One of the most intriguing findings of this analysis was the significantly correlated expression profiles between BRCA1 and the genes corresponding to three DNA damage response pathways. Such expression characteristics suggested that the BRCA1 protein in disease-free group was perhaps still capable of coordinating cellular responses to the apparent DNA damages in breast cancer. Although clear correlation had been established between genetic mutations of BRCA1 and hereditary breast cancer, the precise functional roles of the BRCA1 in cell was not well understood. Based on its interaction with various cellular proteins, it was suggested that BRCA1 played a significant role in a number of cellular processes, including DNA repair, checkpoint control and ubiquitylation and chromatin remodeling[7]. How BRCA1 regulated the functions of different cellular pathways has been of great interest. Our data analysis showed BRCA1 in the disease-free samples was apparently coordinating the communications between different activated pathways, while such coordinative activity of BRCA1 was absent in the metastasis samples. In the disease-free group, we found the expression of BRCA1 was correlated with pathways that were involved in CDC2-induced mitotic arrest, MMR, DNA replication and the HR events that were mediated by BLM (figure 1b). That signaling events coordinated through BRCA1 in the disease-free samples could be an important factor contributing to the viability of these cells. BRCA1 was known to be part of a genome-surveillance complex (BASC), an essential mechanism for maintaining chromosomal stability of the cell[8]. Along with the Nijmegen breakage syndrome 1 (NBS1) and the RAD50-MRE11 complex, the BASC also contained ATM, the MSH2-MSH6 complex, RFC, and BLM. The loss of coordinated regulation by BRCA1 in the metastasis group allowed these cells to evade the regulatory control mechanism. Previous studies showed a decreased BRCA1 expression level in 30% of the invasive breast cancers[9]. It appeared that the effective involvement of BRCA1 in DNA damage response pathways in the disease-free group enabled these cells to maintain their genomic integrity a relatively healthy state.

REFERENCES

1) Zhou, B. B., Elledge, S. J. (2001) The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433-439.
2) Hartwell, L. (1992) Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. *Cell* **71**, 543-546.
3) Wen, L. and Feng, J. A. (2004) RepairFunMap: a functional database of the DNA

repair systems. *Bioinformatics* **20**, 2135-2137.

4) Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.
5) Wood, R. D., Mitchell, M., Sgouros, J., Lindahl, T. (2001) Human DNA repair genes. *Science* **291**, 1284-1289.
6) Van't Veer, L. J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536.
7) Venkitaraman, A. R. (2002) Cancer Susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171-182.
8) Wang, Y. et al. (2000) BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev.* **14**, 927-939.
9) Kennedy, R. D., Quinn, J. E., Mullan, P. B., Johnston, P. G., Harkin, D. P. (2004) The role of BRCA1 in the cellular response to chemotherapy. *J. Nat. Cancer Inst.* **96**, 1659-1668.

IV. SECTION G:

GRADUATE STUDENT POSTERS

GRADUATE STUDENT POSTER ABSTRACTS

G1. Custom design of a multi-application microarray for *Toxoplasma gondii*

Amit Bahl, Manjunatha N. Jagalur*, David Kulp*, and David S. Roos

Department of Biology, Computational Biology Graduate Program, and Penn Genomics Institute, University of Pennsylvania, Philadelphia PA 19143-6018 USA and *Department of Computer Science, University of Massachusetts, 140 Governors Drive, Amherst MA 01003

In recent years, oligonucleotide microarrays have been employed for a wide variety of applications, from gene identification, promoter mapping, and expression profiling, to polymorphism detection, genotyping, and comparative genomic analysis. Taking advantage of the recently-completed reference genome for the protozoan parasite Toxoplasma gondii (14 chromosomes, 65 Mb), we set out to design a single photolithographic oligonucleotide array capable of simultaneously satisfying many competing interests. The global Toxoplasma research community (~200 labs) requires routine, inexpensive expression profiling of ~8000 annotated genes, and tools for genotyping based on well-characterized SNP markers. Individual researchers are also interested in more specialized applications: promoter characterization by chromatin immunoprecipitation, population studies on polymorphic markers, simultaneous profiling of parasite and host cell response genes, etc. Bioinformatics projects include validation of gene models and splicing variants, analysis of noncoding transcripts, and detection of additional polymorphic markers. From the standpoint of array design, it is also of interest to examine alternative protocols for RNA and DNA labelling, probes for transcript detection, and strategies for SNP detection in this small, haploid eukaryotic genome.

The high capacity Affymetrix platform permits inclusion of perfect match probes for whole genome expression profiling and comprehensive genetic mapping, while also leaving sufficient real estate for studies on other research questions of interest. The final design of this custom hybrid chip includes elements for expression, tiling, and mapping arrays, with probe selection and placement optimized for the T. gondii genome. For expression studies, we included 3'-biased perfect match probes for whole genome expression profiling, along with a limited number of gene-specific and nonspecific (surrogate) mismatch probes, enabling comparative analysis of background subtraction algorithms. We have also incorporated probes for all exons on a single chromosome, allowing comparison of different methodologies for RNA labeling and probe set summarization. The exon array also permits pilot-scale analysis of gene models and alternative splicing. In order to study host response to parasite infections, we have included selected host response factor genes from human and mouse, for simultaneous sampling of host and parasite RNA.

Employing a tiling strategy characteristic of resequencing arrays, we have tiled across selected highly variant T. gondii antigenic genes in order to discover strain-specific polymorphisms. Exploratory analysis of murine resequencing arrays has been used to define a tiling resolution that achieves acceptable performance. For transcript discovery, selected unannotated T. gondii ORFs will be tiled at a resolution that yields acceptable results in the simulation. Mapping studies in T. gondii include a set of well characterized genetic markers, and many candidate SNPs discovered through EST analysis of multiple independent parasite isolates. For the well characterized markers, inclusion of multiple probe 'quartets' on each strand permits the use of commercially available software for genetic mapping, and will allow for an empirical determination of probe requirements for genotyping this low complexity haploid genome at the desired sensitivity and specificity. For the additional SNPs, a more limited number of match only probes for each allele should provide cost- and space-effective genotyping at much higher resolution.

G2. Galaxy: A platform for interactive large-scale genome analysis

Dan Blankenberg
Penn State

Galaxy is a metaserver for interactive analysis of genomic data. It is a highly portable, self contained system, complete with a lightweight webserver, embedded database, and a multi-threaded job manager. Galaxy provides a user friendly web portal allowing users to search remote data sources, combine data from independent queries, and visualize the results. At the core of the user's experience is a history system, which stores the actions performed by each individual user. Acting on their history, users are able to undertake various tasks; performing operations such as intersections, unions, and subtractions; and executing other biologically relevant computational tools. Galaxy is designed to handle a variety of data formats, while mitigating the hassles that users face with format incompatibilities. The data types currently supported by Galaxy include genomic interval data (such as that available from the UCSC Table Browser), sequences, and alignments.

Galaxy provides the framework necessary to seamlessly and effortlessly integrate tools. Any command line tool can be added to Galaxy with the creation of an XML configuration file, which specifies the parameters required by the tool, any help to be provided to the user, and the file formats accepted and produced by the tool. Using this configuration file, Galaxy is able to generate all HTML required for the incorporation of a tool.

A number of EMBOSS and embossified PHYLIP tools have been assimilated into Galaxy. Using data obtained from UCSC and other sources, users are able to conduct a pipeline of phylogenic analysis. For example, a user can obtain multiz8way alignments from UCSC (in MAF format), convert this data to FASTA, create a nucleic acid distance matrix, and create neighbor joining trees; all this is accomplished via simple web interface, without any need for programming experience or worrying about file formats. Once trees have been built, a user can then create and download a graphical representation.

G3. A Quantitative Comparative Analysis of Ortholog Identification Methods

Feng Chen, Aaron Mackey and David S. Roos
Department of Biology, Chemistry Graduate Program, and Penn Genomics Institute
University of Pennsylvania, Philadelphia PA 19143-6018 USA

As more and more complete genome sequences become available, the identification of ortholog groups becomes increasingly important, for both functional analysis and comparative evolutionary genomics. This problem is particularly difficult for eukaryotic organisms, due to their large size, extensive fragmentation by introns, and frequent gene duplication and/or fusion. We have previously described a BLAST-based algorithm for ortholog identification, which improves on the well-known COG algorithm in several ways: recent paralogs are identified as reciprocal 'better' hits, a normalization step compensates for consistent between-species biases, and a Markov clustering algorithm is used to define ortholog groups. The entire process operates without requiring manual intervention by expert curators/annotators.

To critically evaluate performance of the OrthoMCL algorithm, we carried out quantitative analysis of a two-species data set (Arabidopsis thaliana, Caenorhabditis elegans), using six alternative ortholog identification methods: RBH (Reciprocal Best Hits), KOG (euKaryotic Ortholog Groups of proteins), INPARANOID, OrthoMCL, RIO (Resampled Inference of Orthology) and Orthostrapper/HOPS. Latent Class Analysis (LCA) was used to define false positive (FP) and false negative (FN) rates of ortholog detection, using maximum likelihood estimation. As expected, RBH (the most widely used approach) shows the lowest FP rate, but its inability to accurately represent many-to-many relationships also yields a high FN rate. Tree-based methods (RIO, Orthostrapper/HOPS) performed similarly, exhibiting an unacceptable FN rate. Performance was better for BLAST-based methods, with KOG and OrthoMCL exhibiting the lowest FN rates, and OrthoMCL and INPARANOID exhibiting substantially lower FP rates than KOG.

For multi-species comparisons, attention was focused on OrthoMCL and KOG. The stand-alone version of OrthoMCL was applied to the KOG reference dataset, containing all protein sequences from seven complete eukaryotic genomes. More than 50% of KOG clusters were grouped identically by OrthoMCL, indicating comparable performance for the automatic analysis provided by OrthoMCL and the manually-curated KOG database. ~35% of the KOG groups were split into smaller groups by OrthoMCL, indicating that KOG groups are more inclusive, but comparison with Enzyme Commission (EC) annotations and protein domain data suggests that OrthoMCL groups exhibit a higher degree of accuracy in both protein function and protein domain architecture.

In order to apply the functional prediction capabilities of OrthoMCL, proteome data from 55 organisms (virtually all complete eukaryotic genomes, and a diverse set of prokaryotes) was clustered into ortholog groups (http://orthomcl.cbil.upenn.edu/). These results were then used to annotate the metabolic pathways of Toxoplasma gondii, a prominent parasite associated with congenital neurological disease and opportunistic infections in immunosuppressed individuals. This analysis provides a surprisingly robust first-draft annotation of metabolic pathways in T. gondii, and comparison with other apicomplexan parasites -- including organisms that cause malaria (Plasmodium sp.), opportunistic infections associated with AIDS (Cryptosporidium sp.), and veterinary disease (Theileria sp.) -- reveals potential targets for drug development, as well as distinctive gene amplifications, losses, and cases of horizontal gene transfer in each taxon.

G4. Clustering of Genes into Regulons using Integrated Modeling (COGRIM)

Gary Guang Chen
CBIL, PCBI, University of Pennsylvania

The computational approaches that are used to identify regulatory modules and networks have traditionally used information either from expression data, sequence features (ChIP binding data or binding motif data) of transcription factors (TF). Although those approaches have been proven useful, their power is inherently limited by the fact that each data resource provides only partial information: expression data provides only functional or indirect evidence, whereas binding data or binding motifs only provide physical location information. Recent efforts on integrating these data types have drawbacks, such as arbitrary parameter cutoffs or too heuristic with little systematic modeling. We present a Bayesian hierarchical model and Markov Chain Monte Carlo implementation that integrates heterogeneous information including expression data, sequence features in a principled and robust fashion. Comparing to available methods, the highlights of our approaches are: (1) the ability to combine two fundamental data types including expression data and sequence features (ChIP-ChIP binding data or binding motif); (2) a framework that allows genes to belong to multiple regulatory clusters, allowing several biological pathways to be modeled simultaneously; (3) no reliance on pre-clustering on expression data and reduced dependence on arbitrary parameter cutoffs; (4) the flexibility to facilitate modular structure discovery and time series data incorporation. By applying our model to geome-wide ChIP binding data and approximately 500 expression experiments on S. cerevisiae, our model successfully captures essential regulatory activities. We found that roughly half of the TF target genes inferred from ChIP binding data are not functional, and that 14% percent of genes that were not considered as TF targets by using binding data alone with stringent p-value threshold were identified by our method as functional target genes. Also, 84 TF pairs were identified to have significant effects on expression of functional target genes. Our validation analyses show that those refined regulons are very likely to be functional TF target genes involved in relevant biological pathways. Our general approach of Bayesian modeling for integrating heterogeneous biological data to discover regulatory networks provides a framework for overcoming the intrinsic limitations of available methods, and should prove useful in applications to other organisms.

G5. Rapid and Asymmetric Divergence of Duplicate Genes in the Human Gene Coexpression Network

WEN-YU CHUNG[1,6], REKA ALBERT[2], ISTVAN ALBERT[5], ANTON NEKRUTENKO[3,5,6] AND KATERYNA D. MAKOVA[4,6*]

Departments of [1]Computer Science and Engineering, [2]Physics, [3]Biochemistry and Molecular Biology, [4]Biology, [5]Huck Institute for Life Sciences, and [6]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, PA, 16802 USA

Abstract

Motivation: While gene duplication is known to be one of the most common mechanisms of genome evolution, the fates of genes after duplication are still being debated. In particular, it is presently unknown whether most duplicate genes preserve (or subdivide) the functions of the parental gene or acquire new functions. One aspect of gene function, that is the expression profile in gene coexpression network, has been largely unexplored for duplicate genes.

Results: Here we build a human gene coexpression network using human tissue-specific microarray data and investigate the divergence of duplicate genes in it. The topology of this network is scale-free. Interestingly, our analysis indicates that duplicate genes rapidly lose shared coexpressed partners: after approximately 50 million years since duplication, the two duplicate genes in a pair have only slightly higher number of shared partners as compared with two random singletons. We also show that duplicate gene pairs quickly acquire new coexpressed partners: the average number of partners for a duplicate gene pair is significantly greater than that for a singleton (the latter number can be used as a proxy of the number of partners for a parental singleton gene before duplication). The divergence in gene expression between two duplicates in a pair occurs asymmetrically: one gene usually has more partners than the other one. The network is resilient to both random and degree-based *in silico* removal of either singletons or duplicate genes. In contrast, the network is especially vulnerable to the removal of highly connected genes when duplicate genes and singletons are considered together. Thus, duplicate genes rapidly diverge in their expression profiles in the network and play similar role in maintaining the network robustness as compared with singletons.

G6. Molecular Evolutionary Analysis of Presynaptic Genes

Dexter D. Hadley[1,2], Tara K. Murphy[2], Otto Valladares[2], Lyle Ungar[1,4], Junhyong Kim[1,4,5], Maja Bucan[1,2,3]

Penn Center for Bioinformatics / Genomics and Computational Biology Graduate Group[1], Department of Genetics[2], School of Medicine[3], Department of Computer & Information Sciences / School of Engineering and Applied Sciences[4], Department of Biology / School or Applied Sciences[5]; University of Pennsylvania, Philadelphia PA 19104

Abstract

To facilitate identification of *cis*-regulatory elements involved in the transcriptional and translational control of gene expression in the neuronal synapse, we initiated a large-scale comparative analysis of genes implicated in presynaptic function. Although annotation of both protein- and non-protein-coding annotation is available through a number of public databases, such datasets are highly automated and somewhat limited in resolution. Thus, we sought to complement these genome-wide efforts by focusing on 130 presynaptic genes (63Mb), and providing highly curated, in-depth annotation of their genomic neighborhoods. Evolutionary analysis combined with bioinformatics approaches, represent a powerful mechanism for understanding the genomic landscape, and we have employed such methods here. By first focusing on regions undergoing purifying selection and then determining various measures of biological importance *in silico*, we prioritize genomic elements for both *in vitro* and *in vivo* verification. In particular, computational approaches include determining the rate of protein evolution, estimating codon bias on coding sequences, and calculating the folding energy of noncoding elements. In this paradigm, we have annotated novel transcripts, missed exons, candidate miRNAs and putative miRNA targets. In addition to considering genes individually, we also consider them in the context of their gene family where applicable. In so doing, we are beginning to explain diverse gene ontologies, gene expressions and other readily available phenotypes from an evolutionary perspective.

G7. Systematic analysis of conservation relations in *E. coli* genome-scale metabolic network reveals novel growth media

Marcin Imielinski[1], Calin Belta[2], Harvey Rubin[3], and Adam Halasz[4]
University of Pennsylvania

The metabolic network is the biochemical machinery with which a cell transforms a limited set of nutrients in its environment into the multitude of molecules required for growth and survival. The advent of sequencing technology and genomic annotation has allowed genome scale metabolic models to be built for many microbial organisms (4).

Current approaches to the study of genome scale metabolic models employ an analysis of feasible and optimal behaviors subject to structural, steady state, thermodynamic, and capacity constraints (4). Structural constraints arise from the stoichiometry matrix, whose columns encode the inputs and outputs of each reaction in the metabolic network. Steady-state constraints follow from the rapid rate of metabolic reactions relative to slower environmental and cellular changes. Thermodynamic (or irreversibility) constraints arise from directionality restrictions on reaction fluxes. Capacity constraints can be derived from the availability of nutrients,the presence of a knockout, biochemical data on the maximum throughput of enzymes, and gene / protein expression data. All of the above constraints restrict feasible flux configurations through the network to a polyhedral set (4). The conservation relations of a metabolic network are linear combinations of species concentration that remain invariant to all flux configurations through the network (1, 6). Represented as vectors in metabolite space, the conservation relations of a metabolic network form the left null space of the stoichiometry matrix. Semipositive conservation relations have been of particular interest because they are associated with the conservation of chemical moieties, atomic elements, and mass (1, 3, 6). The set of semipositive conservation relations associated with a stoichiometry matrix is a polyhedral cone, which can be generated by a unique set of extreme rays, also called extreme semipositive conservation relations (ESCR).

ESCR have the special property being the simplest semipositive conservation relations obeyed by the system, i.e. there exists no semipositive conservation relations obeyed by the network that employs a strict subset of the species of an ESCR. ESCR are closely associated with the distributions of the largest chemical subunits whose structure is preserved by all reactions in a metabolic network (6). ESCR have also been shown to correspond to biologically meaningful metabolite pools (1, 3, 6).

[1] Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA
[2] Dept. of Manufacturing Engineering, Boston University, Brookline, MA
[3] zDept. of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA
[4] GRASP Lab, University of Pennsylvania, Philadelphia, PA

G8. Identification of genes implicated in angiogenesis using an automated text mining process

Jayanti Jagannathan[1], John M. Maris[1,2], Peter S. White[1,2]
[1]Division of Oncology, Children's Hospital of Philadelphia, and [2]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104

Angiogenesis is the process of building of new blood vessels. In a healthy body, this process is required for wound healing and in cancer, it is essential for solid tumor growth. Having as much knowledge about genes implicated as possibly playing a role in the process of angiogenesis is extremely useful in developing agents to target these as a therapy for controlling cancer. The richest source of gene annotation data exists in the biomedical literature and is constantly expanding (>2,000 new articles/month). However, the unstructured nature of biomedical text poses a challenge for the efficient extraction of useful information. Here, we have demonstrated a method based on text mining to automatically identify a set of genes implicated in angiogenesis. A subset of MEDLINE abstracts selected as being relevant for angiogenesis were annotated with a named entity tagger that can automatically identify mentions of genes and proteins. The tagged genes were normalized against an exhaustive human gene lexicon that was built to contain all gene names and all their aliases. The normalized genes were ranked according to frequency of mention. The resulting gene list was validated by comparing it to a list of angiogenes compiled manually by experts, as well as to online resources listing angiogenesis-related genes. The final gene list contained 2,460 unique official gene names. All but 2 of the genes in the manual list were also identified by the text-mined list [99.2% recall (sensitivity)]. The text-mined list was relevance ranked based on the number of documents in which each gene was mentioned, and the 247 highest-ranking genes were compared with the comparably-sized manual list for precision. All of the 100 highest-ranked text-mined genes were identified as being legitimately associated with angiogenesis after further literature review. Furthermore, article recall was 17-fold higher than for articles linked to genes in Entrez Gene, and gene recall was 95-fold higher than for genes assigned to angiogenesis-related GO terms through AMIGO, demonstrating the current under-annotation of these resources for human genes. The method's performance was assessed using a gene expression array-based classifier distinguishing angiogenic from non-angiogenic tumor samples (human). The findings suggest clearly the usefulness of this method in assimilating and annotating genes involved in a biological process from descriptive text.

G9. Identifying and extracting malignancy types in cancer literature

Yang Jin [1], Ryan T. McDonald [2], Kevin Lerman [2], Mark A. Mandel [4], Mark Y. Liberman [2,4], Fernando Pereira [2], R. Scott Winters [3], Peter S. White [1,3,‡]

Departments of [1] Pediatrics and [2] Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia PA 19104 USA, [3] The Children's Hospital of Philadelphia, 34th and Civic Center Blvd. Philadelphia PA 19104 USA, [4] Linguistic Data Consortium, University of Pennsylvania, 3401 Walnut St. Suite 400A, Philadelphia PA 19104

[‡] **To whom correspondence should be addressed.**

ABSTRACT

Summary: MTag is an application for identifying and extracting clinical descriptions of malignancy presented in text. The application uses the machine learning technique Conditional Random Fields and incorporates domain-specific features. MTag was tested with 1,010 training and 432 evaluation documents pertaining to cancer genomics. Our experiments resulted in 0.85 precision, 0.82 recall, and 0.83 F-measure on the evaluation set.
Availability: The software is available at http://bioie.ldc.upenn.edu/index.jsp
Contact: yajin@mail.med.upenn.edu

INTRODUCTION

The biomedical literature collectively represents the acknowledged historical perception of biological and medical concepts, including findings pertaining to cancer research. However, the rapid proliferation of this information makes it increasingly difficult for researchers and clinicians to peruse, query, and synthesize it for biomedical knowledge gain. Automated information extraction methods, which have recently been increasingly concentrated upon biomedical text, can assist in the acquisition and management of this data. Much of this effort has focused upon molecular object (entity) classes, including gene/protein names and protein interactions, and entity recognition algorithms for these tasks have improved considerably in the last few years (Leek 1997, Collier *et al.* 2000, Tanabe and Wilbur 2002, Yu *et al.* 2003, GENIA 2004, Temkin *et al.* 2003, Huang *et al.* 2004). We recently extended this focus to include genomic variations (McDonald *et al.* 2004). Although there have been efforts to apply automated entity recognition to the identification of phenotypic and disease objects (Friedman et al. 1995; Hahn et al., 2000), these systems often do not perform as well as those utilizing more recently evolved machine-learning techniques for such tasks as gene recognition. However, medical entity class recognition is an important prerequisite for utilizing structured text information to improve clinical applications.

To determine the feasibility of efficiently capturing disease descriptions, we describe here an algorithm for automatically recognizing a specific disease entity class: malignant disease labels. This algorithm, MTag, is based upon a Conditional Random Fields model successfully employed in recognizing other biomedical entities (McDonald and Pereira 2004, McDonald *et al.* 2004). The algorithm considers a large number of syntactic and semantic features of the text surrounding each putative mention. MTag directly takes MEDLINE-formatted abstracts from PubMed as input. The output consists of a text file containing a list of identified malignancy types and an HTML file displaying color-coded malignancy types highlighted in the original abstract text. To the best of our knowledge, MTag is the first direct effort at automated literature extraction of a specific disease class. Immediate applications of this algorithm include automation-assisted generation of exhaustive vocabularies and subsequent utility for complex query expansion.

TASK

Our task was to develop an automated method that would accurately identify and extract strings of text corresponding to a clinician's or researcher's reference to cancer (malignancy type). Our definition of the extent of malignant type was generally the full noun phrase encompassing a mention of a cancer subtype, such that "neuroblastoma", "localized neuroblastoma", and "primary extracranial neuroblastoma" were considered to be distinct malignant type mentions. Attached prepositional phrases, such as "cancer <of the lung>", were not allowed, as these constructions often denoted ambiguity as to exact type. Within these confines, the task included identification of all variable descriptions of particular malignant types, such as the forms "squamous cell carcinoma" (histological observation) or "lung cancer" (anatomical location), both of which are underspecified forms of "lung squamous cell carcinoma".

METHOD

In order to train and test the tagger with both depth and breadth, we combined two corpora, for testing. The first concentrated upon a specific malignancy (neuroblastoma) and consisted of 1000 randomly selected abstracts identified by querying PubMed with the query terms "neuroblastoma" and "gene". Of these, 158 abstracts were manually eliminated if they appeared to be non-topical, had no abstract body, or were not written in English. The second corpus consisted of 600 abstracts previously selected as likely containing gene mutation instances for genes commonly mutated in a wide variety of malignancies, and for which genomic and malignant annotations had been previously performed manually. These sets were combined to create a single corpus of 1442 abstracts. This set was manually annotated for tokenization, part-of-speech assignments (Kulick *et al.* 2003, Upenn Biomedical Information Extraction Group, 2004), and malignant type named entity recognition, the latter in strict adherence to our pre-established entity class definition (http://www.cis.upenn.edu/~mamandel/annotators/ent-genrules.html). Dual pass annotations were performed on all documents by experienced annotators with biomedical knowledge, and discrepancies were resolved through forum discussions. A total of 7303 malignant type mentions were identified in the document set.

Based on the manually annotated data, an automatic malignancy type tagger (MTag) was developed using the probability model Conditional Random Fields (CRFs) (Lafferty *et al.*, 2001). We have previously demonstrated that this model yields state-of-the-art accuracy for recognition of biomedical named entity classes (McDonald and Pereira 2004, McDonald *et al.* 2004). CRFs model the conditional probability of a tag sequence given an observation sequence. We denote that $\boldsymbol{O}$ is an observation sequence, or a sequence of tokens in the text, and $\boldsymbol{t}$ is a corresponding tag sequence in which each tag labels the corresponding token with either *Malignancy Type* (meaning that the token is part of a malignancy type mention) or *Other*. CRFs are log-linear models based on a set of feature functions, $f_i(t_j, t_{j-1}, \boldsymbol{O})$, which map predicates on observation/tag-transition pairs to binary values. As shown in the formula below, the function value is 1.0 when the tag sequence is malignancy type; otherwise (o.w.) it is 0. A particular advantage of this model is that it allows the effects of many potentially informative features to be simultaneously weighed. Consider, for example, the following feature:

$$f_i(t_j, t_{j-1}, O) = \begin{cases} 1.0 & \begin{array}{c} t_j = \text{Malignancy Type},\ t_{j-1} = \text{Malignancy Type} \\ O_j = \text{cancer},\ O_{j-1} = \text{lung} \end{array} \\ 0 & O.W. \end{cases}$$

This feature represents the probability of whether the token "cancer" is tagged with label malignant type given the presence of "lung" as the previous token. Features such as this would likely receive a high weight, as they represent informative associations between observation predicates and their corresponding labels. A set of observation predicates, including word and character-*n*-gram characterizations and orthographic predicates (e.g. capitalization patterns) were defined. In addition, we created biomedically-derived predicates, including regular expression patterns (e.g. the suffix –oma) and specified lexicons [e.g. terms from the National Cancer Institute (NCI) neoplasm ontology.] All predicates were then applied over all labels, applying a token window of (-1, 1) to create the final set of features. In total there were six feature types together with 80,294 unique features. The MALLET toolkit (McCallum 2002) was used as the implementation of CRFs to build our model.

RESULTS

Manually annotated texts from the corpus of 1442 MEDLINE abstracts were used to train and evaluate MTag. MTag was tested with a randomly selected 1,010 (70%) training and 432 (30%) evaluation documents pertaining to cancer genomics. The tagger took approximately 6 hours to train on a 733 MHz PowerPC G4 with 1 GB SDRAM Mac server. Once trained, MTag can tag a new abstract in a matter of seconds.

For evaluation purposes, manual annotations were treated as gold-standard files (100% annotation accuracy). The evaluation set of 432 abstracts comprised 2,031 sentences containing malignant type mentions and 3,752 sentences without mentions, as determined by manual assessment of entity content. The predicted malignancy type mention was considered correctly identified if, and only if, the predicted and manually labeled tags were exactly the same in content and both boundary determinations. The performance of MTag was calculated according to the following metrics: Precision (number of entities predicted correctly divided by the total number of entities predicted), Recall (number of entities predicted correctly divided by the total number of entities identified manually), and F-measure ((2*Precision*Recall)/(Precision+Recall)). Our experiments resulted in 0.85 precision, 0.82 recall, and 0.83 F-measure on the evaluation set. Additionally, the two subset corpora (neuroblastoma-specific and gene-specific) were tested separately. The tagger performed with higher accuracy with the more narrowly defined (neuroblastoma) corpus than with the corpus more representative for various malignancies (gene-specific). The neuroblastoma corpus performed with 0.88 precision, 0.87 recall, and 0.88 F-measure, while the gene-specific corpus performed with 0.77 precision, 0.69 recall, and 0.73 F-measure. These results likely reflect the increased challenge of identifying malignant type mentions in a document set demonstrating a more diverse collection of mentions.

Performance of the tagger relative to a baseline system that could be easily employed by a typical research group was also evaluated. For the baseline system the NCI neoplasm ontology, a term list of 5,555 malignant types, was used as a lexicon to identify malignancy type mentions. Lexicon terms were individually queried against text by exact string matching. A subset of 39 abstracts randomly selected from the testing set, which together contained 202 malignancy type mentions, were used to compare the automated tagging and baseline results. The tagger identified 190 of the 202 mentions correctly (94.1%), while the NCI list identified only 85 (42.1%), all of which were also identified by the tagger. Analysis of the results suggested that the major deficiencies of the lexical approach were the inability to identify minor variations in spelling and form (e.g. neuroblastomas), and the inability to identify acronyms (e.g. AML).

MTag has been engineered to directly accept downloaded files from PubMed and formatted in MEDLINE format as input, and to output text and HTML file versions of the tagger results. The text file is similar to the input file, except for the identified malignancy types appended at the end. The HTML file shows the original abstract with color-highlighted malignancy types as demonstrated in the following tagged MEDLINE abstract by Bruder *et al.*:

Normal text
Malignancies

PMID: 15316311
Morphologic and molecular characterization of *renal cell carcinoma* in children and young adults.
A new WHO classification of ***renal cell carcinoma*** has been introduced in 2004. This classification includes the recently described ***renal cell carcinomas*** with the ASPL-TFE3 gene fusion and ***carcinomas*** with a PRCC -TFE3 gene fusion. Collectively, these tumors have been termed Xp11.2 or TFE3 ***translocation carcinomas***, which primarily occur in children and young adults. To further study the characteristics of ***renal cell carcinoma*** in young patients and to determine their genetic background, 41 ***renal cell carcinomas*** of patients younger than 22 years were morphologically and genetically characterized. Loss of heterozygosity analysis of the von Hippel - Lindau gene region and screening for VHL gene mutations by direct sequencing were performed in 20 tumors. TFE3 protein overexpression, which correlates with the presence of a TFE3 gene fusion, was assessed by immunohistochemistry. Applying the new WHO classification for ***renal cell carcinoma***, there were 6 clear cell (15 %), 9 papillary (22 %), 2 chromophobe, and 2 collecting duct ***carcinomas***. Eight ***carcinomas*** showed translocation carcinoma morphology (20 %). One ***carcinoma*** occurred 4 years after a ***neuroblastoma***. Thirteen tumors could not be assigned to types specified by the new WHO classification: 10 were grouped as unclassified (24 %), including a unique ***renal cell carcinoma*** with prominently vacuolated cytoplasm and WT1

expression. Three ***carcinomas*** occurred in combination with ***nephroblastoma***. Molecular analysis revealed deletions at 3p25-26 in one ***translocation carcinoma***, one ***chromophobe renal cell carcinoma***, and one ***papillary renal cell carcinoma***. There were no VHL mutations. Nuclear TFE3 overexpression was detected in 6 ***renal cell carcinomas***, all of which showed areas with voluminous cytoplasm and foci of papillary architecture, consistent with a ***translocation carcinoma*** phenotype. The large proportion of TFE3 " translocation " ***carcinomas*** and "unclassified " ***carcinomas*** in the first two decades of life demonstrates that ***renal cell carcinomas*** in young patients contain genetically and phenotypically distinct tumors with further potential for novel ***renal cell carcinoma*** subtypes. The far lower frequency of ***clear cell carcinomas*** and VHL alterations compared with adults suggests that ***renal cell carcinomas*** in young patients have a unique genetic background.

MTag can be utilized and further explored in various ways. First, when combined with expert evaluation of output, it can help build a vocabulary for all the synonyms of cancer names, which is of great benefit for data integration procedures requiring normalization of malignant types. However, unlike molecular entity classes such as genes, such supervised lists are often not readily available, due in part to the variability in which phenotypic and disease descriptions can be described, and in part to the lack of nomenclature standards in many cases. Secondly, to the best of our knowledge, MTag is the first significant effort to automatically extract entity mentions in a disease-oriented domain. Therefore, this is an important contribution towards a process of identifying and extracting associations between molecular and clinical objects in an automation-centric manner. MTag and its underlying algorithm have been designed to be rapidly adaptable to other biomedical entity classes. Thus, as MTag performs well for extracting malignancy types, this procedure can subsequently be expanded to extract additional disease-oriented information, including clinically-derived observations. Future work will include determining how well similar taggers perform for identifying mentions of malignant attributes with greater (e.g. tumor histology) and lesser (e.g. tumor clinical stage) semantic and syntactic heterogeneity.

ACKNOWLEDGEMENTS

The authors thank members of the University of Pennsylvania Biomedical Information Extraction Group; Kevin Murphy for annotations, discussions and technical assistance; and Richard Wooster for corpus provision. This work was supported in part by NSF grant ITR 0205448.

REFERENCE

Bruder, E., Passera, O., Harms, D., Leuschner, I., Ladanyi, M., Argani, P., Eble, J.N., Struckmann, K., Schraml, P., Moch, H. (2004) Morphologic and molecular characterization of renal cell carcinoma in children and young adults. *American Journal of Surgical Pathology*, 28:1117-1132.

Collier, N., Nobata, C. and Tsujii, J. (2000) Extracting the names of genes and gene products with a hidden Markov model. *In Proceedings of the 18th International Conference on Computational Lingustics (COLING'2000)*, Saarbrucken, Germany.

Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., and Clayton, P.D. (1995) Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1:1-28.

GENIA. (2004) http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Hahn, U., Romacker, M., Schulz, S. (2000) medSynDiKATe: A natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics* 67:63-74.

Huang, M.H., Zhu, X., Hao, Y., Payan, D.G., Qu, K., and Li, M. (2004) Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics,* 20:3604-3612.

Kulick, S., Liberman, M., Palmer, M. and Schein, A. (2003) Shallow semantic annotations of biomedical corora for information extraction. *In Proceedings of the Third Meeting of the Special Interest Group on Text Mining at ISMB 2003.*

Lafferty, J., McCallum, A. and Pereira, F. (2001) Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML-01*, pp. 282-289.

Leek, T.R. (1997) Information Extraction Using Hidden Markov Models. Dissertation, University of California, San Diego.

McCallum, A.K. (2002) http://mallet.cs.umass.edu/

McDonald, R. and Pereira, F. (2004) Identifying gene and protein mentions in text using conditional random fields. In *A Critical Assessment of Text Mining Methods in Molecular Biology Workshop*, 2004.

McDonald, R.T., Winters, R.S., Mandel, M., Jin, Y., White, P.S. and Pereira, F. (2004). An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics,* 22:3249-3251.

Tanabe, L. and Wilbur, W. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics,* 18:1124-1132.

Temkin, J.M. and Glder, M.R. (2003) Extraction of protein interaction information from unstructured text using a content-free grammar. *Bioinformatics*, 19:2046-2053.

Upenn Biomedical Information Extraction Group (2004) http://bioie.ldc.upenn.edu/

Yu, H., Hatzivassiloglou, V., Rzhetsky, A. and Wilbur, W.J. (2002) Automatically identifying gene/protein terms in MEDLINE abstracts. *Journal of Biomedical Informatics,* 35:322-330.

G10. UNTITLED

Michael Gormley
Drexel University

ABSTRACT

Background: Breast cancer is a heterogenous disease with varying clinical outcomes. Histopathological methods of tumor classification have proven to be inconsistent in assessing prognosis. Analysis of cDNA microarray data has shown that gene expression profiling can be used to classify tumors based on the measurement of molecular features of disease. For use of microarrays to gain acceptance in the clinical setting, results must be verified with other methods of quantifying gene expression on large cohorts of patients. Verification becomes problematic due to the limitations of other methods and the size of the datasets involved. Identification of small prognostic gene sets (25-50 genes) would facilitate this process.

Results: A gene expression profile, consisting of 534 genes used to distinguish five molecular subtypes of breast cancer: luminal A, luminal B, ERBB2+, basal, and normal breast-like, was obtained from the Stanford Microarray Database. The gene set was reduced incrementally using significance analysis of microarrays (SAM) to identify and rank the most discriminative genes. Subsets were created consisting of the top 250, 150, 100, 50, and 25 genes and analyzed for their ability to recreate the classifications produced from the original gene set. Performance of the smaller gene subsets varied depending on molecular subtype however; 86% of tumors were classified correctly with as few as 50 genes. Next, gene expression profiles that have shown association with clinical outcome were collected from the literature. Sizes of the expression profiles ranged from 534 to 15 genes. The prognostic power of each expression profile was compared by measuring correlation with estrogen receptor status and tumor grade using the Chi-square test and Kaplan-Meier survival analysis. Five of the six gene sets studied in this manner proved to be prognostic.

Conclusions: This study has demonstrated that microarray-based analysis of a small number of genes can be used to produce classifications with clinical implications in the treatment of breast cancer.

G11. Ethanol Adaptation: A case study in design of experiments, new statistical methods, and functional enrichment

Rishi Khan
Thomas Jefferson University

We present a microarray dataset of ethanol adaptation as a platform for the synthesis of several ideas: (1) The advantages of reference design over dye swaps (2) an estimator of the probability of differential expression: local false discovery rate (3) an application of functional enrichment in Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG).

We compare data taken from dye swaps and reference designs with a novel reference sample. There is significantly less dye bias and less variation in the low signal intensity range in the reference design. While LOWESS correction significantly alters the data in dye swaps, there are very low residuals in the reference design dataset. This suggests, for the most part, the data in the reference design is already normalized.

The local false discovery rate, a statistical metric, has been discussed in the statistical community for the past four years. Estimators for this metric have been successfully implemented in the last year. We prove that the local false discovery rate is the probability that a gene is differentially expressed as opposed to a normal p-value which is the probability of seeing the observed data or more extreme data given that a gene in not differentially expressed. We provide a new estimator of the local false discovery rate and heuristics that detect robust threshold values (i.e. a large change in the metric threshold yields a relatively small change in the number of rejected hypotheses). We use this estimator to determine a threshold in the ethanol adaptation dataset such that the local false discovery rate is 20% and the global false discovery rate is 12%.

We perform functional enrichment on the genes classified as differentially expressed by the local false discovery rate estimator. We present a novel method of annotating the genes on the microarray with functional attributes using the HomoloGene database to obtain homologous genes from rat, mouse, and human. We show the most enriched GO terms and KEGG pathways and discuss some of them in detail.

G12. RepairNET: A Bioinformatic Toolbox for Functional Exploration of DNA Damage Response

Wei Li, Jin-an Feng

Department of Chemistry
Center for Biotechnology
Temple University

ABSTRACT

DNA damage response is one of the essential cellular mechanisms to maintaining genetic integration of the cell. Aberration in the mechanism of DNA damage response often results in cancer. We describe here RepairNET, a protein-protein interaction network associated with the DNA damage response. RepairNET was assembled from the published literature by using a protocol that involved computational data-mining of the MEDLINE and manual curation. This network represents the current knowledge on the intrinsic signaling pathways related to the DNA damage response process. RepairNET currently contain more than 1,200 proteins with over 2,100 functional interactions. A number of web interface tools have been implemented to facilitate a user-friendly environment. The users can navigate through the cellular network associated with the DNA damage response via a Java-based interactive graphical interface. In order to help users explore the functional relationships between the interacting proteins, we have assigned functional domains to the proteins in RepairNET. Based on the protein sequences, a total of 365 unique functional domains were assigned. RepairNET is available online at http://guanyin.chem.temple.edu. It could become an essential resource center for cancer research, providing clues to understanding the functional relationship between proteins in the network, and to building scientific models for the mechanism of DNA damage response and cancer proliferation.

G13. Utilizing Domain Knowledge for Improving Peptide-Allele Binding Prediction

Vasileios Megalooikonomou[1], Despina Kontos[1], Nicholas DeClaris[2,3], Pedro Cano[4]
[1] Computer and Information Sciences Department, Temple University, Philadelphia, USA
[2] Department of Pathology, University of Maryland at Baltimore, Baltimore, MD
[3] Department of Electrical and Computer Engineering, University of Maryland at College Park, MD
[4] Department of Laboratory Medicine, University of Texas, MD Anderson Cancer Center, TX
vasilis@temple.edu, dkontos@temple.edu, declaris@eng.umd.edu, pcano@mdanderson.org

***Abstract-* We developed Radial Basis Function Neural Networks (RBFNN) for studying the problem of allele-peptide binding prediction. We explored utilizing prior domain knowledge in order to optimize allele-peptide binding prediction. We investigated the encoding of inputs of the RBFNN considering chemical properties of amino acids, the discovery of motifs in alleles and the dimensionality reduction based on common motifs discovered. We also explored a number of parameters such as the data set size, unknown-binding data generation, model architecture and training algorithms. Our approach improved the prediction accuracy of peptide-allele binding reaching up to 90% for our best models.**

III. INTRODUCTION

The immune system is composed of many interdependent cell types, organs, and tissues that jointly protect the body from infections (bacterial, parasitic, fungal, or viral) and from the growth of tumor cells. This is considered to be the second most complex body system in humans, after the brain. One of the key players in regulating immune response are the T-cells. In order these to be activated, a very critical stimulus is required, which is generated from peptides bound to Major Histocompatibility Complex (MHC) Class I molecules. The human MHC is known as the Human Leukocyte Antigen (HLA).

Binding of peptides to HLA alleles is necessary for immune reaction, but there is a specific limited number of peptides that can bind to a certain HLA molecule. Our immune system has to discriminate between self and non-self peptides in order to regulate the immune system responses appropriately. Predicting this binding is very important to understanding immunity and the importance of computational analysis in this field is increasing with recent advances in both the fields of biology and computer science. In this study we addressed several issues encountered in knowledge discovery and prediction in allele-peptide binding databases.

Predicting MHC class I to peptide binding is a field that computational methods have been applied in order for certain biological patterns to be investigated [1]. The techniques that have been mostly used are Artificial Neural Networks (ANN), Hidden Markov Models (HMMs), Support Vector Machines (SVM) and binding motifs. A review of how ANNs have been used in these immunology reaction predictions can be found in [2]. Usually, a three layer feed-forward ANN is used that predicts binding peptides for a specific MHC molecule. This network has inputs corresponding to a binary vector representing the peptide amino acids. Each amino acid is encoded by a 20-bit number having "1" at the bit representing the particular amino acid and "0" at the rest of the bits. Hence, a 180 input ANN with one output can be trained using binder and non-binder peptides (represented by 9 amino acids) in the dataset. More applications can be found in [3-4]. HMMs have been utilized to model the profiles of peptide binders and accordingly predict the success of a peptide to MHC molecule binding [5]. SVMs have also been recently used for small datasets [6]. Binding motifs, a method already employed in molecular biology for general sequence alignment and homology searches in databases have been utilized as well for MHC to peptide binding prediction [5].

A particular case of ANNs are Radial Basis Function Neural Networks (RBFNNs) [7]. Typical RBFNNs are constructed by one hidden layer with Radial Basis Function (RBF) units and an output layer with only one linear neuron. In contrast to sigmoid functions, radial basis functions have radial symmetry about a center c in the n-dimensional space where n equals to the number of inputs. The spread σ indicates the selectivity of each neuron [8-9]. The functionality of the RBFNN is based on a distance metric (usually the *Euclidean Norm*) that is computed by the RBF units among the input and the selected center of each RBF unit. In order for this functionality to be effective, the input space has to be represented (encoded) in a way that such a distance computation is meaningful and reflects significant properties of the training samples. In the particular case of predicting peptide-to-allele binding, it is very important that the allele-peptide pair input sequences are encoded in a meaningful manner, in order to reflect binding properties and chemical similarities when the distance metric is calculated in the hidden RBF units.

IV. DATASET

The original data set contained alleles, peptides, and known combinations of bindings showing which alleles bind to which peptides. Confirmed non-binding pairs of alleles and peptides were not available. The amino acid sequences of both the alleles and the peptides were also provided. The data set consisted of 426 unique alleles and 1080 unique peptides. Of these, 86 unique alleles were known to bind to at least one specific peptide in the data set, and 1080 unique peptides were known to bind to at least one specific allele in the data set. We considered 1318 unique allele-peptide combinations (complexes) in the data set that were known to bind. In our original dataset, the binding distribution of alleles and peptides was skewed. In the 1318 binding pairs, a few alleles were present most of the time. In particular, one allele (A*02011) participated in 407 pairs. The peptides were more evenly distributed. There is only one peptide that binds to 7 alleles; most of the peptides bind to only one allele. For the prediction model we used 83 polymorphic allele aminoacid positions. These are positions below 199 since the important part of the molecule is in the first 200 amino acids and any amino acid position after 199 is considered not reliable due to incomplete sequence data for various alleles in our database.

To cope with the unavailability of negative data (i.e. confirmed non-binding pairs) we used two classes, "known" and "unknown" for training and testing. For this reason we needed to generate allele-peptide combinations for which it was not known whether the allele binds to the peptide. We used a number of different ways to generate these "unknown" combinations.

1. **Unknown-StaticRandom**: We created the Cartesian product of all the 86 unique alleles that were known to bind to at least one peptide in the data set and all the 1080 unique peptides that were known to bind to at least one allele in the data set (92,880 pairs). Note that the Cartesian product lists all the combinations of peptides and alleles (binding and non-binding pairs). From this Cartesian product we removed the allele-peptide pairs that have a confirmed binding property (1318 pairs). This resulted in 91,562 actual unknown binding pairs of alleles and peptides. We then randomly sampled the resulting set of unknown binding pairs. Since the known binding pairs were 1318, we created a balanced dataset of both classes by randomly selecting 1318 unknown binding pairs out of the 91,562 available ones. This dataset is called static since the 1318 unknown pairs were selected once and were fixed after the selection (this is in contrast to the dataset in item 2 below). Hence, this dataset was built up from 1318 known binding pairs of allele-peptides (class 1) and 1318 unknown binding pairs of allele-peptides (class 2).
2. **Unknown-DynamicRandom:** In this approach we used again the complete set created by the Cartesian product of all possible combinations (pairs) of peptides and alleles available in our original dataset (92,880 pairs) after removing the ones that have a confirmed binding property (1318 pairs). This dataset was particularly employed for cross-validation experiments with ensemble classifiers. We dynamically selected training datasets by selecting "on-the-fly" the set of 1318 pairs of peptides-alleles for which the binding property was unknown. We randomly picked each time different 1318 unknown pairs. Hence, for each cross-validation experiment, we created a different dataset with the 1318 known binding peptide-allele pairs (class 1) and a set of 1318 unknown binding pairs (class 2) randomly selected each time out of the 91,562 available ones.
3. **Unknown-SimilartoKnown**: This dataset was constructed so that we have the same distribution of alleles in the generated known binding set (class 1) as in the unknown binding set (class 2), in order to take care of the skewed binding distribution of certain alleles in the known binding set. This dataset is called Unknown-SimilartoKnown since there is more similarity between the distributions of unknown and known. Both of them are skewed in the same manner. The unknown data in this case were generated manually using a lengthy process. The resulting dataset consisted of 1318 binding and 1318 unknown binding pairs. More specifically, peptides were grouped according to how many alleles they bind with and an equal number of each peptide grouping was selected in the unknown dataset. In the process of constructing the test dataset, we observed that peptides tend to bind to alleles from the same family (A, B, C). Hence, in order to simulate as much as possible actual unknown (or non-binding) pairs we selected for each peptide alleles from contrasting classes. The distribution of the alleles in both datasets was also preserved, especially for the alleles that seem to dominate the known binding pairs (A*02011 and B*27052).

V. METHODS

A. Encoding of Aminoacid Sequences

A form of domain knowledge in the particular application of allele-peptide binding is the chemical properties of amino acids. We have worked on improving the representation of the amino acid sequences as inputs to the neural network. We have used a hierarchy of the amino acids based on their chemical properties, which is illustrated in Fig. 1. Aminoacids A, V, L, I, P, F, W, and M are non-polar; G, S, T, Y, C, N, and Q are polar, not charged; D and E are polar, negatively charged; and K, R, and H are basic, positively charged. We employed tree node labeling of breadth first search in-order traversal with interleaved binary labels to construct new binary and decimal representations that preserve the chemical distances as much as possible. These representations would be more appropriate for RBFNs since the radial units compute the distances between the input vector and the unit center.

Two different encodings were constructed for the amino acids: a binary and a decimal encoding. In the first encoding approach, the amino acids were coded into a binary code that was represented with 6 bits. In the second approach, a decimal encoding of the binary number corresponding to each amino acid was used representing each amino acid with a single decimal number. These different encodings are shown in Table I. The missing value (for an amino acid) was represented as [111111] in the bit encoding and as the number 38 in the decimal encoding.

B. Basic RBFNN that utilize domain knowledge

In this basic model, the prior knowledge is integrated through the encoding of the amino acids based on their chemical properties (i.e. amino acid hierarchy). The RBFNN had inputs equal to the total number of amino acids representing the peptides and alleles, which is equal to 92. The neurons of the hidden layer utilized the radial basis function for activation whereas the output neuron was linear. In addition to the encoding of the amino acid sequences, we searched the full space of RBFNN optimization parameters to select appropriate values for these parameters and make proper selections of the spread of the Radial Basis Functions (RBFs). The spread was set to 0.5 for all RBFs. The threshold on the prediction error that was used during the training phase was set to 0.01. These models were implemented using the neural network toolbox PRTools 3.1.7 (Pattern Recognition Tools) for Matlab [10].

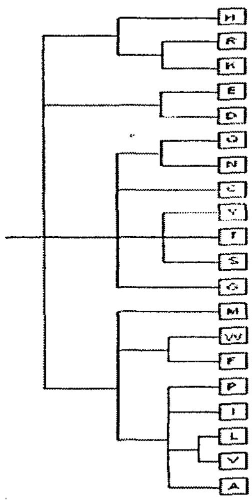


Fig.1. Hierarchical classification of amino acids based on their chemical structure.

TABLE I
THE VARIOUS ENCODINGS USED FOR THE AMINOACIDS

Amino Acids	Bnary Encoding	Decimal Encoding
A	100010	34
C	011100	28
D	110000	48
E	110100	52
F	010000	16
G	100000	32
H	111000	56
I	001100	12
K	111010	58
L	001000	8
M	011000	24
N	111100	60
P	101100	44
Q	111110	62
R	111011	59
S	100100	36
T	100110	38
V	000000	0
W	101010]	42
Y	101000	44
X	111111	38 (median)

C. Motif Discovery – Detecting sequence patterns

In order to further investigate the aminoacid sequence patterns responsible for peptide to allele binding we performed experiments to discover highly conserved aminoacid sequence regions (motifs). For this purpose we used TEIRESIAS [11] (http://cbcsrv.watson.ibm.com/Tspd.html), which is a tool for biological sequence pattern discovery and is distributed freely on the web. The main goal was to identify specific motifs on allele and peptide sequences. The motivation came from the high dimensionality we observed due to the large number of positions we considered for alleles and the insufficient data (initially unbalanced classes) we had to fit the neural network model. Our primary idea was to further reduce the dimensionality of the sequences by removing the positions corresponding to motifs that are present to all the sequences (e.g., reduce from 83 and 9). Another aim was to include the positions of alleles only for specific binding motifs that are present to groups of alleles that bind to specific groups of peptides. We wanted to investigate whether specific binding peptide-allele motifs were present in the dataset. We also intended to use these groupings to reduce the number of different alleles and different peptides we considered. In particular, we planned to explore the possibility of using groups of alleles and peptides that go together to form the centers of the radial basis functions.

We detected motifs present either in all the selected allele sequences (considering only the sequence constructed by the selected and polymorphic positions) or within specific families of alleles (A, B or C). These motifs had high statistical significance and support. We used these findings to reduce the dimensionality of the allele sequences. In the case of the peptides, since the sequences included in our study consisted only of 9 selected aminoacid positions, no significant motifs were detected.

D. Advanced RBFNN models that utilize prior knowledge

In these advances RBFNN models we used again the amino acid encodings based on the hierarchy of their chemical properties. To implement these models we used the Radial Basis Function Toolbox [12] which allows for advance customization of the network parameters and architecture. All the acceptable models we identified used one hidden layer of neurons.

Our network included 433 hidden Gaussian Radial Basis function units with centers selected by the toolbox with a forward selection approach [9]. We also used a regularization parameter λ equal to 10^{-10} in order to penalize large weights during the training process. The error estimation for training was preformed using the generalized cross validation criterion [9] with a threshold of 1/1000 and wait equal to 5 in order to achieve convergence that generalizes well (avoid over-fitting). We performed an exhaustive search through a range of acceptable exponents for the Gaussian functions as well as a range of acceptable radii for the radial units. Trial experiments showed that the best scaling parameter for the specific dataset was equal to 20. Our model had inputs equal to the total number of amino acids used for representing the peptides and alleles, which due to the dimensionality reduction introduced by the allele motif discovered by TEIRESIAS, was equal to 74. One output linear neuron was used to indicate the output label for the test data. A threshold equal to zero was selected as a cutoff value for discriminating between classes labels.

VI. RESULTS

A. Experiments using the basic RBFNN model without utilizing prior knowledge

To provide a basis for comparison for the evaluation of our approach we first performed cross-validation experiments with the Unknown-DynamicRandom dataset and an ensemble of 10 basic RBFNNs classifiers without utilizing any prior knowledge. The amino acids were not coded into any binary form in order to be used as inputs to the classifier model. The PRTools 3.1.7 toolbox for Matlab [10] that we used to implement these initial models has the capability of dealing with letters as inputs.

We performed cross-validation experiments. The spread and the centers of the Radial Basis Units were selected automatically by the training algorithm. The training of the RBFNNs stopped when the error threshold reached 0.02 for the sum-squared error on the training set. The reported accuracy is the average over the accuracies of all 10 classifiers. The classification accuracy obtained ranged between 63% - 72%.

B. Experiments with the basic RBFNN model using the encoding that incorporates chemical properties

In this set of experiments we used again the basic RBFNN model implemented with the PRTools v. 3.1.7 Pattern Recognition toolbox for Matlab [10]. Both the binary and the decimal encoding, based on the chemical properties of the amino acids (i.e. amino acid hierarchy), and were used in these experiments. We used the Unknown-StaticRandom dataset for training and testing. We performed cross-validation experiments varying the percentage of samples used for training. The spread and the centers of the Radial Basis Units were selected automatically by the training algorithm, applying a sum-squared error threshold equal to 0.02 to converge training. Table II illustrates the classification accuracy obtained when using the binary encoding based on chemical properties to represent the amino acids as inputs to the classifiers. The standard deviation in this case ranged between 0.01-0.5. On the other hand, Table II also illustrates the classification accuracy obtained when using the decimal encoding based on chemical properties to represent the amino acids as inputs to the classifiers. The models obtained in this case were much more robust. The standard deviation in this case ranged between 0.01-0.03.

TABLE II

CLASSIFICATION ACCURACY OBTAINED USING RBFNN MODELS THAT UTILIZE PRIOR KNOWLEDGE BASED ON AMINOACID CHEMICAL PROPERTIES

Classification Accuracy with BINARY encoding						
Percent of Data used as training set	**60%**	**55%**	**50%**	**45%**	**40%**	**35%**
Number of Neurons						
5	0.59	0.54	0.48	0.73	0.69	0.49
10	0.59	0.51	0.47	0.73	0.70	0.48
15	0.60	0.54	0.49	0.72	0.70	0.48
20	0.59	0.54	0.48	0,73	0.70	0.48
25	0.60	0.54	0.48	0.72	0.70	0.49
30	0.60	0.55	0.50	0.72	0.69	0.46
35	0.59	0.53	0.49	0.72	0.70	0.49
40	0.59	0.54	0.50	0.72	0.70	0.48
45	0.59	0.54	0.48	0.72	0.69	0.48
50	0.58	0.54	0.50	0.72	0.69	0.48
55	0.59	0.53	0.49	0.72	0.69	0.51
60	0.59	0.54	0.48	0.71	0.69	0.47
65	0.59	0.55	0.49	0.72	0.70	0.48
70	0.60	0.54	0.50	0.72	0.70	0.47
75	0.59	0.54	0.47	0.72	0.70	0.48
80	0.60	0.54	0.50	0.72	0.70	0.48
85	0.60	0.55	0.47	0.73	0.69	0.46
90	0.60	0.53	0.48	0.73	0.69	0.49
95	0.59	0.55	0.50	0.72	0.70	0.49

Classification Accuracy with DECIMAL encoding						
Percent of Data used as training set	**75%**	**70%**	**65%**	**60%**	**55%**	**50%**
Number of Neurons						
5	0.60	0.67	0.70	0.70	0.67	0.55
10	0.60	0.69	0.76	0.73	0.68	0.65
15	0.67	0.70	0.77	0.73	0.71	0.66
20	0.67	0.70	0.77	0.74	0.70	0.67
25	0.67	0.71	0.77	0.74	0.71	0.68
30	0.68	0.71	0.77	0.74	0.71	0.67
35	0.68	0.72	0.78	0.74	0.71	0.68
40	0.68	0.72	0.77	0.75	0.71	0.68
45	0.68	0.71	0.76	0.74	0.71	0.68
50	0.68	0.72	0.78	0.74	0.72	0.68
55	0.68	0.71	0.78	0.76	0.70	0.68
60	0.68	0.71	0.77	0.75	0.71	0.69
65	0.68	0.72	0.78	0.74	0.71	0.68
70	0.69	0.72	0.78	0.75	0.63	0.69
75	0.69	0.72	0.79	0.75	0.72	0.69
80	0.69	0.73	0.78	0.75	0.72	0.69
85	0.69	0.72	0.78	0.75	0.72	0.69
90	0.69	0.72	0.79	0.75	0.71	0.59
95	0.69	0.73	0.78	0.75	0.63	0.69

C. Experiments with advanced RBFNN models that utilize prior knowledge using dimensionality reduction and motif discovery.

To implement these advanced RBFNN models we used the Radial Basis Function Toolbox [12] which allows for advance customization of the network parameters and architecture. The decimal encoding of amino acids, based on their chemical properties, was used. For these experiments we used the Unknown-SimilartoKnown Dataset and motif identification techniques to reduce the dimensionality of the data (i.e., the number of allele positions considered). For motif identification we used TEIRESIAS as explained in the Methods Section. More specifically, in the case of the alleles, we applied TEIRESIAS using the chemical equivalence classes of amino acids provided by the software, removing patterns that overlap. TEIRESIAS requires the setting of several parameters for the detection of motifs, such as the minimum literals in a motif, the maximum window spanned by a motif, and the minimum support of a motif. The particular values selected for these parameters were as follows: L=3 (minimum literals in a motif), W=17 (maximum window spanned by a motif) and K=60 (minimum support). For the model selection the generalized cross-validation (GCV) criterion was used [12]. We detected motifs present either in all the selected allele sequences (considering only the sequence constructed by the selected and polymorphic positions) or within specific families of alleles (A, B or C). These motifs had a high statistical significance and support (e.g. Log-Probability < -30, Occurrences = Sequences = 82). In the experiments reported here, we selected motif present in all the selected allele sequences. By removing this motif from all the peptide-allele sequences the dimensionality of the dataset was reduced to 74, since the motif sequence was 18 aminoacids long. We performed an extensive search for identifying the RBFNN parameters that optimize performance. The classification accuracies obtained for there various experimental settings are illustrated in Tables III to IV.

TABLE III
CLASSIFICATION ACCURACY OBTAINED USING RBFNN MODELS THAT INCORPORATE PRIOR KNOWLEDGE BASED ON AMINOACID CHEMICAL PROPERTIES (DECIMAL ENCODING) AND DIMENSIONALITY REDUCTION BASED ON MOTIF DISCOVERY

Radial Basis Function Radius	Classification Accuracy: Radial Basis Function Exponent										
	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	3.2
1.2	86.44	87.31	87.42	87.96	87.96	88.07	88.18	87.96	88.39	88.50	88.50
1.4	86.88	86.66	87.09	87.53	87.64	87.64	87.85	88.18	88.50	88.18	88.07
1.6	80.15	86.66	86.88	87.20	87.53	87.74	87.64	87.96	87.42	87.53	87.42
1.8	78.63	86.01	86.77	86.77	87.09	87.31	87.42	87.53	87.74	87.42	87.09
2.0	76.25	82.43	86.23	86.55	87.09	86.66	86.98	86.66	86.23	86.66	85.47
2.2	76.90	78.20	85.57	86.01	86.23	87.09	87.53	87.09	87.31	86.88	86.23
2.4	75.92	77.77	85.36	86.23	86.55	85.25	86.01	86.88	86.55	85.90	85.90
2.6	75.38	77.87	85.36	85.03	85.03	85.68	85.03	85.90	85.47	86.23	83.51
2.8	75.27	77.01	81.45	84.16	84.71	85.90	85.68	83.95	84.92	85.57	85.36
3.0	74.73	76.25	78.20	84.38	84.06	85.47	84.60	85.68	85.25	83.84	82.86
3.2	74.19	76.03	78.09	83.95	84.71	85.14	85.03	83.51	85.25	84.27	84.82

TABLE IV
CLASSIFICATION ACCURACY OBTAINED USING RBFNN MODELS THAT INCORPORATE PRIOR KNOWLEDGE BASED ON AMINOACID CHEMICAL PROPERTIES (DECIMAL ENCODING) AND DIMENSIONALITY REDUCTION BASED ON MOTIF DISCOVERY WITH FINALIZED TUNING OF PARAMETERS

Radial Basis Function Exponent	Classification Accuracy: Radial Basis FunctionRadius										
	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8
1.1	88.07	88.07	88.18	88.29	88.18	88.29	88.39	88.29	89.39	89.39	89.29
1.15	87.96	88.29	90.29	88.39	90.07	88.39	88.29	88.39	88.07	90.39	90.61
1.2	88.39	88.39	88.50	88.29	88.50	88.50	88.72	88.50	88.61	88.61	88.50

VII. CONCLUSION

We developed Radial Basis Function Neural Network (RBFNN) models based on Radial Basis Function Units and systematically studied them for the problem of allele-peptide binding prediction. We were able to achieve a prediction accuracy of 90%. Inserting prior knowledge on amino acid chemical properties into the models definitely optimized performance. We clearly demonstrated the proposed digital encoding of the amino acids represents more meaningfully the distance between the amino acids (according to various metrics) and can increase the prediction accuracy. Also, the discovery of motifs in alleles and the dimensionality reduction based on common motifs discovered improved the prediction accuracy of our models. Our attempt to generate "unknown binding" peptide-allele pairs following the distribution of known binding data was quite successful. The dataset "Unknown-SimilartoKnown" demonstrated very reliable accuracy in the neighborhood of 90%. Obtaining actual non-bind data still seems to be very critical in generating even more robust predictive models. Motif discovery which we used as dimensionality reduction improved performance significantly. There is still significant work that can be performed in this area.

We believe we made novel and important contributions in the methodology used for incorporating existing domain knowledge into the prediction models. In particular, the encoding of inputs of the RBFNN considering the chemical properties of amino acids, Incorporation of additional domain knowledge is expected to improve further the prediction accuracy. In addition, we expect that performance will improve by increasing the size and quality of the data set, and by obtaining even a small set of non-bind ("negative") data.

REFERENCES

[1] V. Brusic, N. Petrovsky, G. Zhang, V. Bajic, "Prediction of promiscuous peptides that bind HLA class I molecules", *Immunology and Cell Biology*, 80: 280-285, 2002.

[2] V.Brusic, J.Zeleznikow, "Artificial Neural Networks in Immunology", in Proceedings of the *1999 International Joint Conference on Neural Networks IJCNN'99*.

[3] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network", *Bioinformatics*, 14, 121-130, 1998.

[4] K. Gulukota, J. Sidney, A. Sette, C. DeLisi, "Two complementary methods for predicting peptides binding major histocompatibility complex molecules" *Journal of Molecular Biology*, 26, 1258-1267, 1997.

[5] H. Mamitsuka, "Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models", *Proteins* 33, 460-474, 1998.

[6] P. Dönnes, A. Elofsson, " Prediction of MHC class I binding peptides, using SVMHC", *BMC Bioinformatics* 3:25, 2002.

[7] S. Haykin, *Neural Networks: A comprehensive foundation*. Upper Saddle River, NJ, USA: Prentice Hall, 1999.

[8] M.J.L. Orr. Introduction to radial basis function networks. Technical report, Institute for Adaptive and Neural Computation, Division of Informatics, Centre for Cognitive Science, University of Edinburgh, Scotland, April 1996. www.anc.ed.ac.uk/~mjo/papers/intro.ps.

[9] M.J.L. Orr, "Regularization in the selection of radial basis function centers", *Neural Computation*, 7(3) 606-623, 1995.

[10] R.P.W. Duin, A Matlab Toolbox for Pattern Recognition, Delft University of Technology, The Netherlands, 2002, www.ph.tn.tudelft.nl/prtools/.

[11] I. Rigoutsos, A. Floratos, "Combinatorial Pattern Discovery in Biological Sequences: the TEIRESIAS Algorithm", *Bioinformatics*, 14(1), 1998.

[12] Matlab Functions for RBF Networks, www.anc.ed.ac.uk/~mjo/rbf.html.

G14. Improving Protein Secondary-Structure Prediction by Predicting Ends of Secondary-Structure Segments

Uros Midic[1], A. Keith Dunker[2], Zoran Obradovic[1*]

[1]Center for Information Science and Technology, Temple University, 1805 N. Broad St., 303 Wachman Hall, Philadelphia, PA 19129 USA

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 714 North Senate Avenue, Suite 250, Indianapolis, IN 46202 USA

Abstract – Motivated by known preferences for certain amino acids in positions around α-helices, we developed neural network-based predictors of both N and C α-helix ends, which achieved about 88% accuracy. We applied a similar approach for predicting the ends of three types of secondary structure segments. The predictors for the ends of H, E and C segments were then used to create input for protein secondary-structure prediction. By incorporating this new type of input, we significantly improved the basic one-stage predictor of protein secondary structure in terms of both per-residue (Q_3) accuracy (+0.8%) and segment overlap (SOV_3) measure (+1.4).

G15. Combinatorial Transcriptional Regulation: United We Bind

Sarita Nair
University of the Sciences in Philadelphia

Developing deeper understanding of the regulation at various levels in biological systems will enable us gain useful insights into the basis of differences between various cell types, cellular response to different environmental stimuli, and contribution of signaling pathways to various cellular processes. Eukaryotic transcriptional regulation involves coordination of multiple transcription factors and hence is combinatorial in nature. Combinatorial regulation addressed in this study is based on the concept of composite elements (CE). Composite regulatory elements contain two closely situated binding sites for two distinct transcription factors. Specific factor-DNA and factor-factor interactions contribute to the function of CEs. Coordinated action of transcription factors binding to the CEs results in highly specific patterns of transcription that cannot be individually produced by the involved factors. A database of known CEs is available (TRANSCOMPEL: Kel-Margoulis et al., 2002).

We have developed an approach to study the combinatorial aspect of gene regulation occurring through composite regulatory elements. The key aspect of our research centers on the statistical enrichment analysis of known composite elements in the upstream regulatory regions of co-expressed genes. The workflow includes automatic retrieval of promoter regions of clusters of co-expressed genes, computational identification of known composite elements, and generation of hypotheses on those elements, factors and genes that are playing a role in the biological process under study. The current implementation enables the identification of composite elements using CATCH, a web-based search tool associated with the TRANSCOMPEL database. The statistical significance of composite elements in clusters of co-expressed genes is conducted through Fisher's Exact Test based on the hyper geometric probability of observed number CEs compared to those in a reference list of genes. The reference typically is the micro array from which the gene expression clusters are obtained. Sensitivity analysis is performed by varying the binding site similarity score, distance between the binding sites in the CE and the number of allowed mismatches to the known binding sites. The individual parameter values range from 0.7-1.0 for the similarity score, 0-5 for the inter-site distance and 0-3 for allowed mismatches. Only those CEs that are significantly enriched for all the parameter values are considered in the subsequent generation of hypotheses.

Cross species homology is used at this stage to provide a filter for biologically significant predictions. Comparison of genomic sequences of multiple organisms can provide useful insights into the relevance of statistically enriched composite element binding sites in regulatory sequences. Several resources like VISTA, UCSC genome browser etc provide a comprehensive platform to investigate the evolutionary forces behind functionally important elements. The conservation track of UCSC genome browser is employed to carry out the phylogenetic analysis.

Conservation track is generated using genome wide multiple sequence alignment of eight organisms. These alignments are used to determine "highly conserved elements", using a phylogenetic hidden markov model (phylo-HMM). The phylo-HMM categorizes each alignment of a multiple sequence alignment to conserved or not conserved models. This is achieved by assigning a score to each base; corresponding to its probability of being conserved. Base by base position scores are further used to generate the scores for a conserved element. Current implementation involves checking the enriched composite elements against the identified highly conserved elements to determine the extent of conservation.

We demonstrate our approach in a case study involving dynamic time profiles of differentially expressed genes from the renal proximal epithelial cells (RPTEC) exposed to Staphylococcal Enterotoxin B (Ionin et al., 2005). The approach will be integrated in to PAINT suite for transcriptional regulatory analysis (Vadigepalli et al., 2003).

REFERENCES

Ionin B, Das R, Pontzer C, Jett M. Staphylococcal enterotoxin B induces cytoskeletal rearrangement and apoptosis in human kidney cells. In review.

Kel-Margoulis, O.V., Kel, A.V. et al. TRANSCompel: a database on composite regulatory elements in the eukaryotic genes. Nucleic Acids Research. 2002, 30(1), 332-334.

Panda, S., Antoc, M.P. et al. Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock. Cell. 2002, 109, 307-320.

Siepel, A., Bejerano, G. et al. Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. Genome Research. 2005, 15, 1034-1050.

Vadigepalli, R., Chakravarthula, P., Zak, D.E., Schwaber, J.S., Gonye. GE, PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. Omics. 2003, 7(3), 235-52.

G16. Using Phylogenetic Reconstruction to detect Lateral Gene Transfer events in the apicomplexan parasite *Toxoplasma gondii.*

Lucia Peixoto, Feng Chen, David S.Roos

Department of Biology, and Penn Genomics Institute
University of Pennsylvania, Philadelphia PA 19143-6018 USA

The Apicomplexa are a monophyletic group composed almost entirely of parasitic species, including many important pathogens, including as those responsible for malaria and toxoplasmosis. Genome data for several apicomplexan parasites, including Toxoplasma gondii, have recently become available, making it possible to exploit comparative genomic tools, together with phylogenetic analysis, to provide insights into the adaptation to the parasitic life style and potential targets for therapeutic development. Lateral gene transfer (LGT) is now accepted as a major evolutionary force in prokaryotes, but its contribution to the evolution of eukaryotic organisms has been controversial. There is evidence that LGT has been an important process in Apicomplexan evolution, however, most notably in the acquisition of a secondary endosimbiotic plastid -- the apicoplast -- acquired when an ancestral protist 'ate' a eukaryotic alga, and retained the algal plastid as an essential organelle.

In order to detect candidate instances of LGT in the T.gondii genome, we first established a benchmark of 36 verified LGT events (many associated with the apicoplast). OrthoMCL-DB, which incorporates the complete genomes of 55 organisms spanning the tree of life (http://orthomcl.cbil.upenn.edu/), was used to identify ortholog groups for phylogenetic reconstruction. The ortholog dataset was constrained to specify only those groups including 3-100 sequences, at least one T. gondii sequence, and at least one bacterial, archeal and/or plastid-containing organism (plants/algae), as these represent the most likely sources of LGT. Each group was aligned using MUSCLE, and phylogeny reconstruction was performed using Maximum Likelihood (PHYML), with 100 bootstrap replicates. A reference species tree was defined based on gene content data for the 55 taxa present in this database, and the phylogeny of each ortholog group was compared to the reference using Horizstory, which employs a recursive consolidation and rearrangement procedure to identify incompatible tree topologies and specify probable sources of LGT.

Using a bootstrap cutoff of 80% to consolidate nodes in each tree, 1517 groups were found to present incompatible tree topologies and at least one probable LGT path containing T.gondii. This set includes all 36 proteins in the benchmark dataset, and the correct LGT path was predicted in 35/36 cases. It may seem surprising to consider that 15% of the T.gondii genome could have arisen through LGT events, and other methods for assessing tree incompatibility (NNI metric, Robinson-Foulds distance, etc) will be used for further analysis.. Regardless, the groups identified have already served to highlight several promising targets for drug and/or vaccine development.

G17. Computational Analysis of Gene Regulatory Networks in Retinal Pigment Epithelial Cells during Epithelial-Mesenchymal Transformations

C.H. Pratt, R. Vadigepalli, G.E. Gonye, G.B. Grunwald.

Department of Pathology, Anatomy and Cell Biology, Jefferson Medical College, Thomas Jefferson University, 1020 Locust Street, Philadelphia, PA 19107.

Epithelial-mesenchymal transformation (EMT) is a mechanism of tissue remodeling during which cells transition, either unidirectionally or reversibly, between two behavioral states comprised of populations of either cohesive cells forming solid tissues (epithelium), or solitary, migrating cells (mesenchyme). EMT occurs as a normal component of many biological processes including embryonic development and wound healing, such as during gastrulation of the early embryo, or repair of epithelial tears. In addition, abnormal EMT occurs during pathobiological processes such as metastasis and invasion of cancer cells from a primary tumor site. Thus, EMT plays a fundamental role in many biological processes, and much has been learned regarding the underlying cellular and molecular mechanisms governing these cell behavioral changes. However, since EMT involves the integration of numerous cellular subsystems and coordinated expression of a large number of genes and proteins, it will be of interest to identify those higher level processes, such as gene regulatory networks, that orchestrate these events.

To begin to investigate this, we have utilized the Retinal Pigment Epithelium (RPE) as a model system. The RPE is a simple monolayer epithelium located at the back of the eye, whose interactions with the neighboring neural retina are important for the development and maintenance of a healthy functional visual system. The RPE has been shown to undergo both a normal EMT in response to a wound in the epithelial sheet, and an aberrant response known as Proliferative Vitreoretinopathy (PVR) after rhegmatogenous retinal detachment (including a torn retina), which can result in a loss of normal vision. We hypothesize that PVR results from an aberrant wound healing process whereby the normal balance of EMT-related behaviors is shifted towards the mesenchymal phenotype, due to inappropriate cell signaling and resultant gene expression, resulting in a fibrotic response and retinal scarring. The purpose of this study was to begin to elucidate a candidate global regulatory network involved in the reciprocal regulation of genes during epithelial-mesenchymal transformations of Retinal Pigment Epithelial cells, with the long-term goal of identifying potential therapeutic targets.

In this study, we initially selected a set of sixty-one genes whose expression is altered in RPE cells during EMT, including specific markers of the respective mesenchymal and epithelial cells state in RPE cells, such as the adhesion molecules N- and R-cadherin. Using the analysis tool PAINT (Promoter Analysis and Interaction Network Toolset; www.dbi.tju.edu), cognate promoter sequences were retrieved for each gene. These promoter sequences were analyzed to identify transcription regulatory elements (TREs) within the sequences. A statistical analysis of the data, performed by PAINT, identified the TREs that were over-represented and under represented ($p > 0.1$) in our selected gene list as compared to our reference gene list. PAINT then integrated the statistical analysis, gene list and identified TREs into a Candidate Interaction Matrix. This analysis was repeated for homologous genes found in the human, mouse and rat genome. Promoter models for reciprocally regulated genes and for N- and R-cadherin were then constructed using those TREs that were over-represented ($p > 0.1$) and were present on at least one promoter sequence in two of the three species. Our analysis identified individual TREs that were enriched among the promoters analyzed, suggesting that these transcription factors may be important in coordinating RPE gene expression. These included Nkx2-5, E2F1, IRF-7, Poly A, Oct-1, Barbie Box, Sox-5 and FoxJ2 elucidated as potential global regulators of EMT and FoxD3, HNF3α, HNF1, TFII-I, E2F1/DP-2, SREBP-1, AML-1, Pax and COMP1 found as potential specific regulators of N- and R-cadherin. RT-PCR was then used to identify those candidate transcription factors whose mRNA is actually detectable in RPE, and furthermore those which were differentially expressed in either cultured (i.e. mesenchymal, N-cadherin-expressing) or fresh (i.e. epithelial, R-cadherin-expressing) RPE tissue, respectively. This computational analysis has identified a specific subset of TREs that may be of importance in regulation of RPE EMT in general, and for cadherin regulation in particular. Supported by grants NIH RO1EY06658, NIHT32ES07282, and DARPA F30602-01-2-0578.

G18. Phenotypic Heterogeneity of Breast Cancer Tumoroids

A. Vamvakidou, P. I. Lelkes, A.Tozeren

Drexel University, School of Biomedical Engineering, Science & Health Systems,
3141 Chestnut Street, Philadelphia, PA 19104
av54@drexel.edu

Cancer is a multi-step process leading to uncontrolled cellular proliferation, tumor formation, and possible metastasis as disease progression occurs. Human cancer cell lines establish effective model systems for *in vitro* analysis of tumorigenesis. In a recent stochastic model of tumor growth developed in our laboratory, tumors are classified as aggressive A1 (low metastatic, hormone dependent), A2 (increased proliferation and migration) and A3 (highly metastatic, hormone independent, drug resistant) based on prognostic factors such as tumor type, size and grade.

The aim of our research is to experimentally establish growth, migration, and cell-cell interaction data of cultured and cocultured breast cancer cell lines for inclusion in a computational model of tumor growth. We have selected three breast cancer cell lines MCF-7 (A1), ZR-75-1 (A2) and MDA-MB-231 (A3) which are representative of each of the above phenotypes and compare specific tumor cell traits, such as proliferation, metabolism, migration and tumoroid formation in Rotating Wall Vessel (RWV) bioreactors.

CellTracker™ probes (Molecular Probes) were used to fluorescently label with different colors the three cancer cell lines. Cell migration assays were performed using 24-well inserts (3-μm pore size; BD Biosciences). Cells that migrated through the membrane and attached to the bottom were fixed and stained. Membranes were cut out and photographed, and migratory cells were counted. Growth rates and morphology of tumoroids grown in RWVs were characterized over time by paraffin embedding and haematoxylin and eosin staining of harvested samples. Histology was also done on frozen tissue sections to preserve the fluorescence of the cells in coculture.

Preliminary experiments measuring cellular metabolism and proliferation with the Alamar Blue™ assay (Biosource) indicate that metabolic levels, and consequently proliferation, are increased in more "aggressive" breast cancer cells. Interestingly, cocultures of moderately invasive and minimally invasive cell lines resulted in the reduction of the overall level of cellular metabolism, implying a dynamic cell-cell interaction.

G19. Overlapping Reading Frames in Eukaryotes

SAMIR WADHAWAN, PAULA GOETTING-MINESKY, KATERYNA MAKOVA AND ANTON NEKRUTENKO

Penn State University

GNAS1 gene is one of the most extensively studied examples, where two structurally unrelated mammalian proteins, *XLas* and *ALEX* are translated from alternative overlapping reading frames of a single gene. Not only are they encoded by the same locus, but a specific *XLas/ALEX* interaction is essential for G-protein signaling in neuroendocrine cells. A disruption of this interaction leads to abnormal human phenotypes including mental retardation and growth deficiency. The region of overlap between the two reading frames evolves at a remarkable speed: the divergence between human and mouse *ALEX* polypeptides makes them virtually unalignable! We have shown that the two proteins are locked in an evolutionary race: in New World monkeys and Old World monkeys. *ALEX* accumulates the bulk of amino acid replacements whereas in human and apes *XLas* takes the lead. *XLas/ALEX* is the first example of a rapidly evolving locus encoding interacting proteins via overlapping reading frames with a possible link to the origin of species-specific neurological differences. This prompted us to ask if there were other such cases in eukaryotic genomes. Preliminary analysis of the yeast genome suggests a positive answer to this question.

G20. Microarray Analysis In Macrophages Infected With Suicidal Transgenic *Leishmania Spp*

Hsiuan-Lin Wu, Malik Yousef, Michael Nebozhyn, Celia Chang and Louise Showe

The Wistar Institute

Leishmania spp. are uniquely suitable among trypanosomes to serve as a live vaccine models because of the following characters: 1. naturally infect antigen-presenting cells i.e. dendritic cells and macrophages; 2. amenable to laboratory maintenance in culture and in animal models; 3. non-pathogenic; 4. amenable to genetic manipulations for expressing endogenous or foreign genes allowing the accumulation of the foreign proteins for efficient antigen presentation. One principal disadvantage common to all live vaccine models is the issue of residual pathogenicity. The development of a vaccine model in which the vaccine carrier could be efficiently eliminated would be very useful. In this study we examined the effects on gene expression in a mouse macrophage cell line infected with a transgenic *Leishmania* mutated in the heme biosynthesis pathway that could be induced to commit suicide by exposure to delta-aminolevulinate (ALA) a pathway substrate and/or UV. Intra-macrophage *Leishmania* amastigotes are lysed selectively when rendered uroporphyric *in situ* by treatment with ALA or UV with little detectable effect on the host. The model tested was to compare gene expression patterns in a mouse macrophage cell line infected with the transgenic Leishmania, or mock infected with a control plasmid. The cells were then treated with ALA and/or UV to induce Leishmania death.

The investigation of effects on macrophage gene expression in response to infection and in response to additional treatments with ALA and/or UV for selective suicidal *Leishmania* cytolysis has been carried out using microarrays. A total of 28 RNA samples, corresponding to four different experimental conditions for both infected and mock-infected macrophages (untreated, ALA, UV, ALA+UV), each done in quadruplicate, were analyzed on cDNA microarrays carrying ~38,000 mouse clones and probes for 4 *Leishmania* genes. Identification of significant genes was performed using t-test and ANOVA with subsequent filtering of genes that had relative fold change less than 2-fold. The number of differentially expressed genes detected as being statistically significant substantially exceeded the estimated number of false positives, indicating reliable detection of systematic changes in the gene expression levels caused by the infection and the ALA/UV treatments. In addition, the expression levels of the Leishmania genes were monitored to determine efficiency of infection and subsequent killing upon treatment with ALA/UV. Significant changes were detected in many genes related to the immune response after infection including the induction of some immuno-suppressive genes. After ALA and UV treatments to induce *Leishmania* cytolysis, many genes that had been significantly affected by the infection, reverted back to their normal expression levels. This suggests that the removal of the parasite allows normal macrophage function to resume suggesting the transgenic porphyric *Lieshmania* may be a useful model for a live vaccine.

V. SECTION P:

POSTER ABSTRACTS

POSTER ABSTRACTS

P1. A modular analysis of a stem cells populations compendium.

Ghislain Bidaut and Christian J. Stoeckert Jr.

Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA
ghbidaut@pcbi.upenn.edu

Stem cells hold great promise in tissue regeneration medicine. Understanding stem cell differentiation and self-renewal mechanisms, as well as elucidation of cellular mechanisms defining the state of "stemness", is crucial to the effective use of stem cells. To this end, we created a compendium dataset of gene expression data measured in a variety of experiments spanning various stem cells populations including hematopoietic SCs, liver, prostate, bladder, kidney and bone, in mouse, human, and zebrafish. This data was generated by the SCGAP (Stem Cell Genome Anatomy Projects – http://www.scgap.org) consortium participants.

For the comparison of gene expression in different tissues in a global fashion, all experiments were normalized in terms of expression call and annotations: Stem cells populations were annotated with a controlled vocabulary describing stem cell differentiation stages (multipotent, totipotent, progenitors, lineage-commited progenitors, differentiated cells). A heat map was generated for visualization and clustering, allowing the study of the compendium using standard microarray analysis algorithms. To overcome the problem of heterogeneous data, we measured the enrichment of (KEGG) pathways and created a map of significantly enriched cellular processes across all the populations.

Preliminary results shows conserved modules shared by all stem cells populations as well as modules specific to a restricted set of tissues. The stem cells compendium data has the potential to help the elucidation of the role and interactions of signaling processes involved in SC differentiation and self-renewal. such as Wnt and Notch signaling pathways, as well as of the transcriptional activities of Otc4 and Nanog factors. Future plans include the creation of a network interaction maps as well as extending the compendium with other stem cells populations.

P2. Genome-wide inference of protein function in *Plasmodium falciparum* using a computational model of the parasite interactome

Shailesh V. Date & Christian J. Stoeckert, Jr.
Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104.

The genome of the human malarial parasite *Plasmodium falciparum* encodes thousands of unknown proteins. Functional characterization of such proteins is of utmost importance, if we are to design combative strategies against malaria based on the biology of the parasite. To infer protein function on a genome-wide scale, we computationally modeled the *P. falciparum* interactome, by integrating experimental and computational functional genomics data within a Bayesian framework. The resulting functional interaction network covers ~68% of the parasite proteome, and allows characterization of more than 2000 unknown proteins, based on their associations with other known proteins. Network reconstruction involved the use of novel strategy, where we incorporated a set of continuously updated, uniform reference priors in our model, given the absence of any prior knowledge about the parasite interactome. This strategy can be generalized, and applied to investigate genomes of other organisms with sparse interaction data. Further, we also superimposed the *P. falciparum* interaction network on genomes of three apicomplexan pathogens- *Plasmodium yoelii*, *Toxoplasma gondii* and *Cryptosporidium parvum*, describing relationships between these organisms based on retained functional linkages. Our comparisons reveal the highly evolved nature of *P. falciparum*; for instance, a deficit of nearly 26% in terms of predicted interactions is observed against *P. yoelii*, due to missing ortholog partners in pairs of functionally linked proteins.

P3. Cooperation among Histone H2B-directed Ubiquitin Protease 10 (Ubp10) and Silent Information Regulator 2 (Sir2) at *S. cerevisiae* Silent Chromatin

N.C. Tolga EMRE[1], Don A. BALDWIN[2], Shelley L. BERGER[1]

1. **Gene Expression and Regulation Program, The Wistar Institute, Philadelphia, Pennsylvania 19104, USA.**
2. **2. Penn Microarray Facility, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.**

The physiological state of the eukaryotic genomes consists of chromatin, where DNA is complexed with nuclear proteins, such as histones. Chromatin structure provides the necessary organization and compaction to the genomes, while modulation of this structure, such as through post-translational chemical modifications of histones, is important for a variety of nuclear processes, including gene expression and silencing.

In the budding yeast *S. cerevisiae*, the expression of genes proximal to chromosome ends (telomeres) is epigenetically silenced similarly to positioning close to heterochromatin in diverse organisms like human cells and *Drosophila*. Among other histone modifications, ubiquitylation of histone H2B, methylation of histone H3, and acetylation of histone H4 are important for silent chromatin in the budding yeast. It has been postulated that low levels of histone methylation and acetylation within silent chromatin, and high levels within euchromatin, help to sequester silent information regulatory (Sir) proteins to silenced regions due to Sir proteins' binding affinity to under-modified histone tails. Recently, we have shown that the ubiquitin specific protease Ubp10/Dot4, previously linked to silent chromatin, targets histone H2B for deubiquitylation both *in vivo* and *in vitro*. We have further shown that Ubp10 maintains low H2B ubiquitylation and is instrumental in keeping H3 Lys4 and Lys79 methylation levels low proximal to the telomere, which helps the proper localization of the silencing factor Sir2 histone deacetylase to this region.

We have also shown that proper localization of Sir2 and Ubp10 to silent telomere-proximal regions are mutually dependent on each other. In order to investigate a possible cooperation among these factors for their silencing functions, we have employed genetic and genomic approaches. First, we show that a yeast strain lacking *UBP10* and containing a hypomorphic allele of *SIR2* show more severe defects in silencing of a telomere-proximal transgene than either mutations alone. We then extended our analysis to look at genome-wide expression profiles of these mutants. We observe that, at several chromosome ends, combined *SIR2* and *UBP10* mutations have an increased defect in silencing of endogenous telomere-proximal genes, in terms of magnitude and/or regions affected, compared to either mutation alone, suggesting that under wild type conditions, both Ubp10 and Sir2 are required to define the chromosomal regions to be silenced. These observations are consistent with a view that Ubp10 and Sir proteins cooperate in establishing the silent chromatin near chromosome ends. We are planning to extend our studies using genome-wide localization experiments to investigate how the levels and distribution of the histone modification marks mentioned above are affected at the silent chromatin in the mutant yeast strains.

P4. Mouse PromoterChip BCBC-5A: a powerful tool for genome-wide location analysis

Regina K. Gorski, Athanasios Arsenlis, Jonathan Schug, Peter White, John E. Brestelli, Phillip Phuc Le, Christian J. Stoeckert, Jr., Klaus H. Kaestner

Genome-wide location analysis can provide insight into transcriptional regulatory networks by identifying genes bound by transcription factors *in vivo*. The Mouse PromoterChip BCBC-5A provides a powerful and low-cost new tool that can be used for location analysis of specific binding by transcription factors as well as global histone modification. The Mouse PromoterChip BCBC-5A contains over 18,000 proximal and distal promoter regions. To identify the promoter regions, transcriptional start sites (TSS) were determined by the BLAT alignment of RefSeqs and full length cDNA libraries to the mouse genome. PCR primers were designed to amplify promoter sequences between -3000 and +200 bp relative to the TSS. Over 12,000 well characterized genes were chosen and are represented by either 1 Kb or 2 Kb tiles, PCR amplified from genomic DNA. Future plans include a second chip that will contain 18,000 additional elements representing enhancers, microRNA genes, highly conserved genomic regions and additional promoter regions. Genome-wide location analysis with the promoter chip can be combined with expression profiling using the Mouse PancChip. The two microarray platforms have been developed in concert and contain a large number of common genes, allowing for orthogonal analysis of genes regulated in the mouse. Publications resulting from the use of the Mouse PromoterChip BCBC include *Friedman et al.* Proc. Natl. Acad. Sci. 2004, *Le et al.* PLoS Genetics 2005, and *Rubins et al.* Mol. Cell Biol. 2005. Annotation and other information about the promoter chip are available at http://www.cbil.upenn.edu/EPConDB/.

P5. EXPLORING BIOCHEMICAL NETWORKS WITH REACHABILITY ANALYSIS

Ádám M. Halász[1], Saurav Pathak[1], Marcin Imieliński[1,2], Oleg Sokolsky[1], Mark Goulian[3], Harvey Rubin[2] and Vijay Kumar[1]
[1]GRASP Lab, [2]School of Medicine, and [3]Department of Physics,
University of Pennsylvania, Philadelphia, PA 19104-6228

Introduction

With the increase in experimental data and an increasing interest in modeling it makes sense to formulate biologically relevant queries as reachability questions, for at least two reasons. On one hand, the corresponding dynamical systems are typically high dimensional and have nonlinear interactions, making analytical and intuitive insights difficult. On the other hand, the quantitative data is subject to significant uncertainties, due to both natural variability and experimental limitations.

Reachability allows us to delineate the reachable states from a given set of plausible initial states, or to determine the set of plausible initial states that lead to a set of final states. In this presentation we discuss two recent examples where reachability analysis can yield relevant biological insight.

Our method (Belta et al., 2002), relies on dividing the state-space into hyper-rectangles and replacing the nonlinear functions describing the kinetics with different multi-affine approximations in each hyper-rectangle. The linear interpolation property of multi-affine functions allows a computationally efficient procedure to identify necessary conditions for reachability between hyper-rectangular sets in state space.

First we discuss the extension of our earlier work on the lactose metabolism in *E.coli* to a more complete model which includes the inhibiting effect of glucose. The lactose-glucose metabolism of E.coli is particularly interesting because it exhibits bistability, as seen in the phenomenon of induction of the lac operon. During exponential growth, the bacterium can process lactose at either a high or at a low rate, for a range of external lactose concentrations. Induction or switching to the high rate occurs when a threshold value of lactose is exceeded. A mathematical model (Yildirim and Mackey, 2003) successfully reproduced this feature. A more complete model (Santillán and Mackey, 2004) takes into account the effect of glucose. Lactose is actually an alternative source of energy to the bacterium, secondary to glucose. When the latter is present, the lactose metabolism is kept in its low state via catabolite repression (enhancement of lac transcription via a signaling molecule, cAMP) and inducer exclusion (blocking of lactose influx by glucose). Catabolite repression is described mathematically by a sophisticated stochastic model of the lac promoter and its interactions with the lac repressor, allolactose, and the cAMP-CRP complex. We show how reachability analysis helps identify the ranges of lactose and glucose concentrations where bi-stability is present.

While the results on the previous example refer essentially to steady state properties, in our second example we illustrate the unique power or reachability analysis to delineate transient features that occur during non-equilibrium time evolution, without relying on direct ODE simulations.

Reachability and Piecewise Affine Systems

Multi-affine functions are affine (of the form $f(x)=ax+b$) in each of their variables, and can be represented as a sum of monomials containing at most the first power of each variable. They have a linear interpolation property, over hyper-rectangles (Cartesian products of intervals of each of their variables). The value of the function at any point inside a hyper-rectangle can be written as a convex combination of the values taken on the vertices of the hyper-rectangle.

Consider a dynamical system described by a state vector X, and a (matrix) velocity function $f(X)$:

$$\dot{X} = f(X) \tag{1}$$

Each element $g(x_1,\dots x_k)$ of $f(X)$ is a multivariate function. We choose a convenient partition of the space of X into hyper-rectangles and replace the original function

Each element $g(x_1,\dots x_k)$ of $f(X)$ is a multivariate function. We choose a convenient partition of the space of X into hyper-rectangles and replace the original function with a multi-affine function on each hyper-rectangle, so that the resulting approximating function is continuous. After doing this for each entry in $f(X)$, we merge the partitions resulting from the individual functions. Each hyper-rectangle in the joint partition has a complete set $f_H(X)$ of multi-affine functions $g_H(x_1,\dots x_k)$.

Suppose a trajectory crosses the common facet between two adjacent hyper-rectangles, *A* and *B*, in the direction of B The velocity vector at the crossing point points towards the interior of *B*, ie, its component orthogonal to the facet must have the appropriate sign. The orthogonal component is a multi-affine function on the common facet, and its value inside the facet is bounded by its values on the vertices (because of the interpolation property). It can have a given sign somewhere on the facet if and only if it also does so on one of the vertices.

This reduces the problem of checking reachability between two adjacent hyper-rectangles to checking the sign of a multi-affine function at 2^d points. Since the vertex is shared by just as many hyper-rectangles, the reachability verification between adjacent hyper-rectangles requires on average one function evaluation. Using this criterion we identify the existence of crossing trajectories in either direction between all pairs of adjacent hyper-rectangles, resulting in a graph whose vertices represent hyper-rectangles, with directed edges between pairs of hyper-rectangles for which crossing trajectories exist. The existence of a path between two nodes is a necessary condition for the existence of a trajectory linking the two corresponding hyper-rectangles.

Since we exploit a necessary condition reachability analysis using this procedure is conservative. Because of this it is possible that the configuration of hyper-rectangles resulting from the piecewise approximation may be too coarse for a given reachability question. In practice, it is useful to subdivide these rectangles to provide additional granularity.

If the dependence on parameters is approximated with piecewise affine functions, the parameters can be treated as variables whose time derivative is zero. One may perform reachability analyses for intervals of parameter values, particularly useful when parameter uncertainty is an issue.

The Glucose-Lactose System

A diagram of the model network is given in Figure 1. The lac operon codes for permease and beta-galactosidase. Permease brings external lactose into the cell and beta-galactosidase converts it into allolactose. Allolactose blocks the lac repressor which otherwise inhibits transcription of the lac operon. In the absence of glucose, the signaling substance cAMP is produced. cAMP binds to CRP to form a complex, cAMP·CRP, which enhances lac transcription. Glucose also inhibits the inbound transport of lactose.

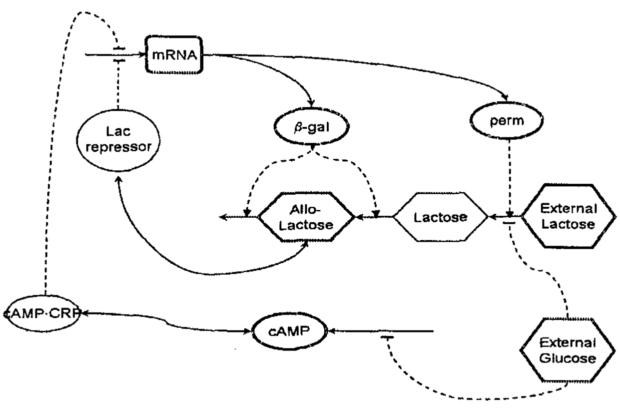


Figure 1: Diagram of the lactose-glucose network. The external glucose and external lactose are the external inputs while the dynamics determine the concentrations of the other quantities.

We follow closely the model by (Santillán and Mackey, 2004), referring the reader there for further details and references. The reactions of cAMP with CRP are assumed to be fast and the concentrations of cAMP·CRP and of free cAMP are algebraic functions of the total concentration of cAMP. The activity of the lac promoter is described by a thermodynamical model which gives the transcription rate as a function of the concentrations of complex and (free) lac repressor. The dependence of the free lac repressor on the concentration of allolactose is also given algebraically. The amount of allolactose is set to one half of that of lactose inside the cell. The above dependencies are summarized by two functions, $\eta([A_t],[cAMP_t])$ for the activity of the lac promoter and $\omega([cAMP_t])$ for the concentration of free cAMP. The dynamical part of the model describes the evolution of six variables representing the concentrations of β-galactosidase, mRNA, permease mRNA, β-galactosidase, permease, total allolactose, and total cAMP, $\{[M_b],[M_p],[B],[P],[A_t],[cAMP_t]\}$:

$$
\begin{aligned}
[\dot{M}_b] &= [P1]k_m\eta([cAMP_t],[A_t])_{\tau_B} - (\mu+\xi_M)[M_b] \\
[\dot{M}_p] &= [P1]k_m\eta([cAMP_t],[A_t])_{\tau_P} - (\mu+\xi_M)[M_p] \\
[\dot{B}] &= \tfrac{1}{4}\kappa_B[M_b]_{T_B} - (\mu+\xi_B)[B] \\
[\dot{P}] &= \kappa_P[M_p]_{T_P} - (\mu+\xi_P)[P] \\
[\dot{A}_t] &= \tfrac{1}{2}\varphi_{L_1}[P]\left(\frac{[L_E]}{\Phi_{L_1}+[L_E]}\frac{\Phi_{G_1}}{\Phi_{G_1}+[G_E]}\frac{[A_t]}{[A_t]+\Phi_{L_1}/2}\right) \\
&\quad -\tfrac{1}{2}\varphi_{L_2}[B]\left(\frac{[A_t]}{[A_t]+\Phi_{L_2}/2}\right) \\
[c\dot{AMP}_t] &= \varphi_C\frac{\Phi_C}{\Phi_C+[G_E]} - \xi_C\,\omega([cAMP_t]) - \mu[cAMP_t].
\end{aligned}
\tag{2}
$$

The concentrations of external glucose $[G_E]$ and lactose $[L_E]$ are inputs. The remaining symbols are constants taken from (Santillán and Mackey, 2004). The first four equations contain time delays which we ignore as they do not modify the steady state structure and we do not expect them to significantly impact the reachability properties.

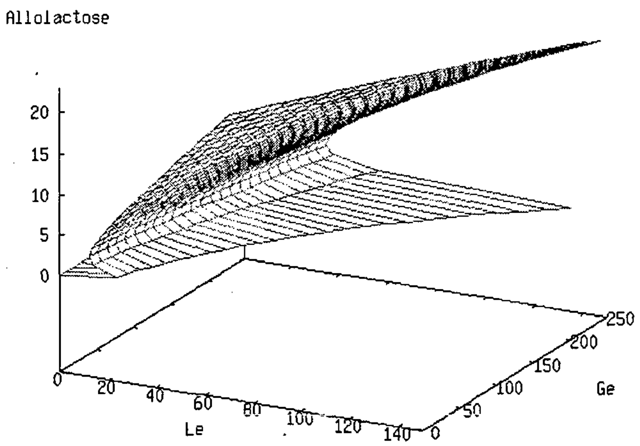


Figure 2: Steady state values of allolactose as a function of external glucose and lactose. All units are µM.

The steady states of the system (2) can be calculated by setting the equations of motion to zero. A diagram of the steady state values of allolactose is given in Figure 2. The surface is 'folded' for a section of the lactose-glucose plane, found between two threshold values of Le (Figure 4). For these parameter values there are three solutions to the steady-state equations, of which only the outer two are stable (as shown numerically).

We are interested in the possibility of induction, when the system is caused to evolve into a high internal lactose state. It can be achieved by temporarily increasing lactose, decreasing glucose or a combination of the two.

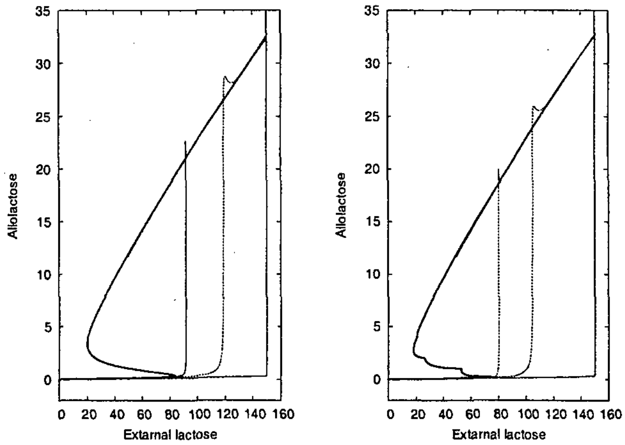


Figure 3: Steady state values and simulation traces for allolactose, in the exact (left) and in the piecewise linear model (right). All units are µM.

Figure 3 shows several upward switching trajectories obtained by gradually increasing Le over a time T at fixed Ge, as well as the corresponding steady state curves. The trajectories turn upward for some value of Le *exceeding* the location of the kink in the steady state curve. If the increase of Le is fast, the upturn occurs for Le values significantly higher than the threshold, so this method is not reliable in predicting bistability and induction.

We expect that the likely 'inducible' region in the Le-Ge plane is the one with only the high steady state, to the right of the curve corresponding to the edge of the fold in the steady state surface (Figure 4). This intuition needs to be verified using calculations of Jacobians and eigenvalues in a six dimensional state-space. Figure 2 only shows one of the relevant dimensions of the system. The regions of attraction of the two steady states can have a complicated geometry, so there is no *mathematical guarantee* that the external lactose and glucose concentrations cannot be manipulated to achieve induction, even without leaving the bi-stable region. Reachability analysis can in fact delineate the region from which an increase in external lactose may lead to induction.

Reachability Analysis

The first step in our procedure is the piecewise multi-affine approximation of the equations of motion (2). The functions that need to be approximated are $\eta([A_t],[cAMP_t])$ and $\omega([cAMP_t])$ and the nonlinear terms involving $[A_t]$. We defined a grid of values for the two variables and computed tables of values The approximating functions are (bi-)linear interpolations between the tabulated values. The steady states and trajectories for this system closely reproduce the exact ones, as illustrated in Figure 3.

This linearization procedure only partitioned the ranges of two model variables. In order to gain insight from reachability calculations, we need to have partitions that delineate regions of interest for the analysis. The partition we used in the calculations below had $12^4 \times 22 \times 9$ hyper-rectangles. It occupies a hyper-rectangular region spanning from zero to 2-10 times the typical unstable steady state value, for all variables, except for cAMP, where we chose intervals to cover the range of its steady state values for the glucose concentrations we considered.

We performed reachability calculations using this grid, for various values of external glucose and lactose. In each calculation we evaluated the set of all hyper-rectangles reachable from the hyper-rectangle including the steady state value of cAMP for the given external glucose and the lowest ranges for all other variables.

Our partition does not contain the high steady state. If the reached set does not include the boundary of the partition then we can conclude that the high steady state is not reachable from the given initial set of configurations hence induction is forbidden.

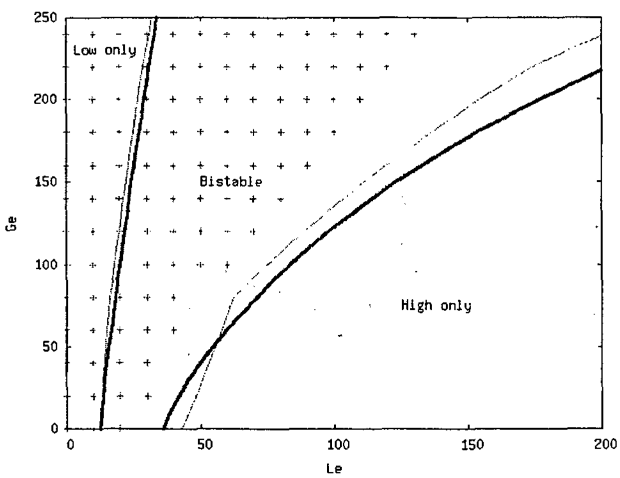


Figure 4: Bistability region in the Le-Ge plane, in the exact model and the piecewise approximation. Ssuperimposed are points where reachability forbids upward switching. All units are µM.

Figure 4 summarizes the reachability results. Points signify values for which induction is not possible according to reachability. The plot also shows the boundary of the bi-stable region. The non-inducible points are inside the bistability region, as expected intuitively.

A similar calculation can be performed with the external lactose and glucose dependence also piecewise approximated. These calculations delineate a set of hyper-rectangles from which it is *impossible* for the cell to evolve to a high internal lactose concentration, *regardless of the external lactose and glucose concentration.*

Tetracycline Resistance in *E.coli*

One of the important means of gram-negative bacterium *E.coli* resistance to the antibiotic tetracycline (*Tc*) is active efflux of the antibiotic from the cell membrane. This is achieved predominantly via the expression of the *tetA* gene in transposon Tn*10*, which produces an inner membrane protein *TetA* (Hillen, 1994). The *TetA* protein binds with the tetracycline in the cell and exchanges it with a proton in the periplasm, thus reducing the antibiotic concentration. The *TetA* protein, which affects the cell membrane electrostatic potential is harmful to the cell and is very tightly regulated by the *TetR* protein. The repressor protein, *TetR* is expressed by the *tetR* gene, represses both tetA and tetR by binding to two proximal *tetO1* and *tetO2*. Incoming tetracycline chelated tetracycline, $MgTc^+$ binds to TetR which weakens its affinity to the operators tetO1 and tetO2, enabling the *tet* genes to express (Lederer, 1995).

The action of the TetA protein is further supplemented by the action of the cell membrane, which influences the internal tetracycline concentration (Thanassi, 1995). We attempt here to model the membrane effect as well as the gene expressions that describe the composite cell response to the tetracycline influx. A simple model is given by a four dimensional system of ordinary differential equation. This system yields transient concentrations that are important for the cell survival. Most of the parameters in the model are not known and are not sufficiently constrained by experimental results. We employ reachability techniques to examine parameter ranges.

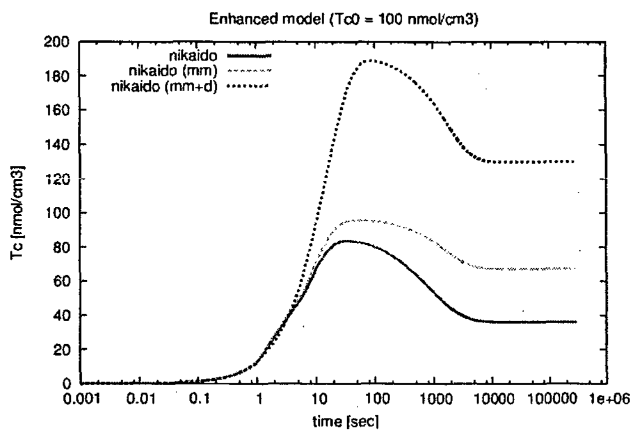
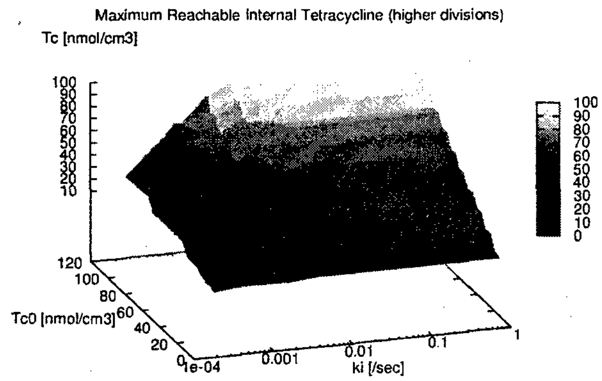


Figure5: System trajectories for several parameter values, illustrating the evolution of internal tetracycline (Tc) concentrations after exposure to a given external concentration.

As illustrated in Figure 5, the internal *Tc* concentration in a typical situation would start from zero, then increase to a transient maximal value. As the defense mechanism is triggered, *TetA* pumps *Tc* out from the cell, gradually reducing its internal concentration which eventually settles at an equilibrium value. Both the steady-state value and the transient maximum are potentially relevant to the survival of the cell. While the former can be calculated from steady state considerations, the knowledge of the latter requires performing a complete dynamical simulation.



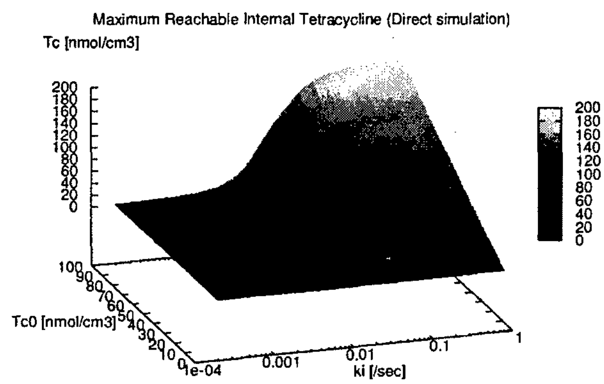


Figure 6: Direct computation (bottom) and reachability calculation (top) of the highest transient Tc concentration. The reachability calculation is overly conservative, but it is much more economical.

Since the model parameters are not well known, our first task is to identify those parameter values that are consistent with other observations (such as lethality of a given *Tc* concentration). One important observable is the maximum transient concentration as well as the time the system spends in the high concentration range. Given this low dimensional model, the direct calculation is also feasible. While Figure 6 shows some significant discrepancies, the general features of the dependence are reproduced.

Conclusions

We applied reachability analysis to a model of the glucose-lactose metabolism of E.coli in order to investigate the possibility of induction under different combinations of glucose and lactose availability. The reachability analysis delineates (a) sets of hyper rectangles that lead to the induced state; and (b) sets of hyper rectangles from which it is not possible to evolve to the induced steady state. These results (exact for a piecewise linear approximation of the model) are obtained without performing extensive simulations or a detailed stability analysis.

In the case of tetracycline resistance, both reachability and traditional methods are applicable. Our examples illustrate how reachability analysis can be used in conjunction with other methods. We used traditional techniques for steady states and simulations. However, numerical simulations only provide snapshots of system behavior. Based on these simulations we can formulate global hypotheses which can be verified by reachability calculations.

Acknowledgments

This work was supported in part by DARPA IPTO through grant no. AF F30602-01-2-0563 . ÁMH is supported by an NIH-NLM Individual Bioinformatics Fellowship.

References

Belta, C., L. Habets, and V. Kumar. 2002. Control of multi-affine systems on rectangles with applications to hybrid biomolecular networks. *In 41st IEEE Conference on Decision and Control.* Las Vegas, NV.

Hillen, W., C. Berens, Mechanisms underlying expression of Tn10 encoded tetracycline resistance. Annu Rev Microbiol, vol 48, 1994 345-69

Lederer, T., M. Takahashi, W. Hillen, Thermodynamic analysis of tetracycline-mediated induction of Tet repressor by a quantitative methylation protection assay. Anal Biochem, vol 232, 1995, 190-6

Santillán, M., M. C. Mackey (2004). Influence of Catabolite Repression and Inducer Exclusion on the Bistable Behavior of the *lac* Operon. *Biophys. J., 86,* 1282.

Thanassi, D. G., G. S. Suh, H. Nikaido, Role of outer membrane barrier in efflux-mediated tetracycline resistance of Escherichia coli. J Bacteriol, vol 177, 1995, 998-10

Yildirim, N., M. C. Mackey (2003). Feedback Regulation in the Lactose Operon: A Mathematical Modeling Study and Comparison with Experimental Data. *Biophys. J., 84,* 2841.

P6. Linking the VIP/VPAC2 signaling model to a circadian clock model predicted VIP phase shifting effect.

Haiping Hao*,#, Daniel E. Zak*,#, Thomas Sauter#,§, James Schwaber#, and Babatunde A. Ogunnaike*

*** Department of Chemical Engineering and Delaware Biotechnology Institute, University of Delaware, Newark, DE 19176**
 # Daniel Baugh Institute for Functional Genomics and Computational Biology, Thomas Jefferson University, Philadelphia, PA 19107
§ Institute for System Dynamics and Control Engineering, University of Stuttgart, Germany

Increasing evidence suggests an important role for VIP/$VPAC_2$ activated signal transduction pathways in maintaining a synchronized biological clock in the suprachiasmatic nucleus (SCN). We have developed a kinetic model integrating $VPAC_2$ activated cAMP/PKA signaling leading to induced circadian clock gene expression and linked the model to a published circadian clock model. Simulation of the linked model matched to VIP induced phase shift both *in vitro* and *in vivo*. Simulations using varied receptor levels also agreed with photoperiod phase shifting data from transgenic animals and predicted single pulse of light or VIP applied to VPAC2 over-expressing transgenic animals will produce similar effect, i.e., more pronounced phase advance. Future works will include: 1). Mapping the VIP/VPAC2 signaling networks in the SCN cells; 2). Mapping the VIP/VPAC2 induced gene expression network; 3). Incorporating additional signal transduction pathways and gene expression network into the model; and 3). Measuring kinetic data for all signaling pathways to obtain better estimate of the parameters.

P7. An approach to identify regulatory modules for tissue-specific transcripts sharing a tissue-specific Gene Ontology Biological Process.

Elisabetta Manduchi, Jonathan Schug, Christian J. Stoeckert Jr.

University of Pennsylvania

Abstract: We are interested in exploring how mammalian promoters specify expression in a given tissue at a certain time for a particular purpose. Our goal is to deconstruct promoters into their primary regulatory components and understand the usage of these components. We have utilized recent human and mouse tissue survey microarray datasets (e.g. Novartis GEA 2) to discover putative regulatory modules characterizing genes preferentially expressed in a given tissue and also sharing a biological process that is highly correlated with that tissue. Here a regulatory module may be described by expressions of the form "binding sites for transcription factors A, B and C within 300 pairs of each other", etc. For a given tissue survey, we utilize a Shannon entropy based method [1] to attach to each gene representative a score for each of the surveyed tissues. This score reflects both the gene's overall tissue specificity (i.e. how much its expression pattern differs from ubiquitous uniform expression) and its categorical specificity, i.e. its specificity to that particular tissue. We then utilize ranking tests based on these scores to identify Gene Ontology (GO) Biological Processes that are significantly specific for a given tissue. Examples of these are steroid metabolism for liver and muscle development for skeletal muscle. For such a pair (tissue, GO biological process) we construct a suitable positive training set and negative training set of promoter sequences, which are then input into a grammar-based approach [2] to identify discriminating modules. We have refined this approach by exploiting human/mouse consistency to establish a final set of potentially relevant grammar productions (each representing a regulatory module), which we subsequently use to build a classifier using random forests.

References:

1. Schug J. et al., Genome Biol. 2005; 6(4):R33. Epub 2005 Mar 29.
2. Schug J., Ph.D. Dissertation, 2005

P8. A bioinformatics approach to identify recoding events of A-to-I RNA editing

Stefan Maas[1], Daniel Lopresti[2], Derek Drake[3], Rikhi Kaushal[1], Stephen Hookway[2], Walter Scheirer[2], Mark Strohmaier[2], and Christopher Wojciechowski[2]

[1] Department of Biological Sciences, Lehigh University, Bethlehem, PA
[2] Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA
[3] Department of Computer Science, Purdue University, West Lafayette, IN

Abstract

RNA editing by A-to-I modification in pre-mRNAs constitutes a major mechanism for the generation of RNA and protein diversity in mammals and is known to regulate important functional properties of neurotransmitter receptors in the central nervous system. A-to-I RNA editing can also create or destroy pre-mRNA splice signals or lead to alterations in RNA secondary structures. We have recently identified widespread editing in 5'-and 3'-untranslated mRNAs involving Alu repeat elements in the human transcriptome (Athanasiadis et al., PLoS Biology 2004) using a combined bioinformatics and experimental screening and validation strategy.

In all known and characterized cases of recoding by A-to-I editing, the ensuing amino acid substitutions have been linked to alterations in protein function. A few additional cases of recoding due to RNA editing were recently identified through bioinformatics approaches but the total number of targets is still low despite evidence that many more should exist.

Here we present our ongoing work to develop a comprehensive screening process to identify site-selective A-to-I editing events in human mRNAs. Our computational screen is implemented as a pipeline involving of a number of processing stages. The current code consists of a set of programs written in the C language running on a Unix-compatible operating system, and makes use of a number of standard built-in utilities and open source software, including the mySQL database server.

At the first stage, the locations of all discrepancies of a specified type between a genomic DNA sequence and its corresponding mRNAs is output. Database tables from the UCSC Genome Browser (http://genome.ucsc.edu/) are downloaded locally and queried using the mySQL server to produce all mRNAs associated with a given chromosome. This list is

then filtered, leaving only the sites for which the discrepancy between the genomic and mRNA sequence reflects an A→G substitution (including proper handling of indeterminate nucleotides) and for which the 'A' is located in a marked exon.

In the next stage, we filter the list of differences recorded so far to remove known SNPs. To be conservative about not excluding potential RNA editing sites that have been mistakenly annotated as SNPs, we can limit the filter to exclude, for example, only SNPs that are labeled in the database as having been derived through genomic-level analysis.

Finally, each potential editing site that remains is individually scored by determining a 50-nucleotide window around it and another 50-nucleotide window within 2,500 bases upstream or downstream for which a foldback score is computed. This score is determined by weighing the matches between different base pairings as follows: C↔G = 3, A↔U = 2, and G↔U = 1. Only the optimal score for all possible window "slides" is saved.

Our approach is being optimized to identify genes that harbor single or few editing events, such as the well characterized codon changes in glutamate and serotonin receptor transcripts. Additional parameterized stages for scoring candidate sequences are being developed and genes that receive the highest overall scores will then be validated in the laboratory for evidence of RNA editing in vivo and assayed for potential downstream effects on the function of the gene products.

P9. Identification of Gene Targets against Dormant Phase *Mycobacterium tuberculosis* Infections

Dennis J Murphy[1] and James R. Brown[2], Bioinformatics, Genetics Research, GlaxoSmithKline. 1250 South Collegeville Road, UP1345. P.O. Box 5089 Collegeville PA 19426-0989.

E-Mail: [1]Dennis.6.Murphy@sgk.com; [2]James.R.Brown@gsk.com

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), kills more than 2 million people per year and has infected an estimated 2 billion people worldwide. Upon infection *M. tuberculosis* is engulfed into the macrophage phagosome. Once inside it is able to attenuate host attack and alter its metabolism to survive in this hypoxic and nutrient deprived environment. Persistent infection due to this dormant state can last months to decades and the lack of current targets against dormancy is the main obstacle in the development of effective therapies against *M. tuberculosis.*

Here we present a meta-analysis of published data on *M. tuberculosis* gene expression under simulated dormancy conditions using DNA microarrays as well as *M. tuberculosis* gene essentiality based on saturation transposon mutagenesis. The goal of our study is to derive a prioritized list of potential gene targets effective against *M. tuberculosis* in the dormant phase. Six published studies monitored expression *in vitro* using hypoxia ('Wayne' model) or starvation to model dormancy. Three *In vivo* experiments followed expression of MTB that survived in mouse macrophages, lung, and subcutaneous hollow fibers. Transposon insertion mutants have been profiled for the ability to grow on Petri plates, in mouse macrophages, and in live mice. We scored each of the ~4000 genes in the *M. tuberculosis* genome based upon two criteria: i) the relevance of the experimental conditions to the persistent/dormant state, and ii) the relative level of expression (in the case of the microarrays) or growth inhibition (in the case of the knockouts). Scores for each gene in over a dozen experimental conditions were collected into a relational database that allowed the data to be parsed and analyzed in several ways.

Preliminary analyses revealed that for the top 5% of upregulated genes in the expression studies, a greater number of genes for virulence and detoxification were highly expressed under *in vivo* compared to *in vitro* conditions. Several chaperonins were among the top expressing genes. The *DOS* regulon, a two component regulatory system, has been previously proposed to be critical for the dormant state. This regulon consists of 43-77 genes, of which, about 60% of these genes are in the top 5% of expressed genes, confirming the importance of this regulatory element in dormancy under a variety of experimental conditions. Our analysis also support proposals in the literature that *M. tuberculosis* reduces protein synthesis while spending its limited resources on maintaining cell wall integrity and membrane potential, and resisting host defenses. For example, a significant fraction of the <u>down</u>-regulated genes include 30S and 50S ribosomal proteins, as well as ATPases utilized in respiration with oxygen. Nearly one quarter of the set of top scoring genes in the expression experiments also result in attenuated growth after transposon mutagenesis, and several are likely essential for growth.

Future work will integrate information on pathways and pharmaceutical tractability in order to derive a new class of anti-TB drug targets.

P10. Robust Knowledge Extraction over Large Unstructured Biomedical Text Collections

Min Song
Center for Information Science & Technology
Temple University

Knowledge extraction is an emerging research area where various research fields need to be incorporated. This multidisciplinary nature of the field has led this project to address and investigate the research problems in the context of assisting the biomedical knowledge workers to find the desired information in an intelligent manner. In this project, we propose RIKE, Robust Iterative Knowledge Extraction, a hybrid knowledge extraction algorithm drawn from several research fields such as pseudo-relevance feedback, data mining, natural language processing, and information extraction. Specifically, we develop a robust novel extraction algorithm that consists of 1) a keyphrase-based query expansion to spot the promising documents and 2) a Mixture Hidden Markov model-based information extraction. We also conduct a series of experiments to validate three research hypotheses formed in this project.

As illustrated in Figure 1, RIKE consists of two major components: DocSpotter, querying and retrieving promising documents for extraction, and HMMIE, a statistical generative model based IE. RIKE queries a search engine with ranked keyphrases, discovers prediction rules from the retrieved unstructured documents, and then the rules are used to predict additional information to extract from future documents, thereby improving the recall of IE.

The underlining algorithm is as follows:

1. Starting with a set of user-provided seed instances (the seed instance can be quite small); our system retrieves a sample of documents from the text databases. At the initial stage of the overall document retrieval process, we have no information about the documents that might be useful for the goal of extraction. The only information we require about the target relation is a set of user-provided seed instances, including the specification of the relation attributes to be used for document retrieval. We construct some simple queries by using the attribute values of the initial seed instances to extract the document samples of a pre-defined size using the search engine.

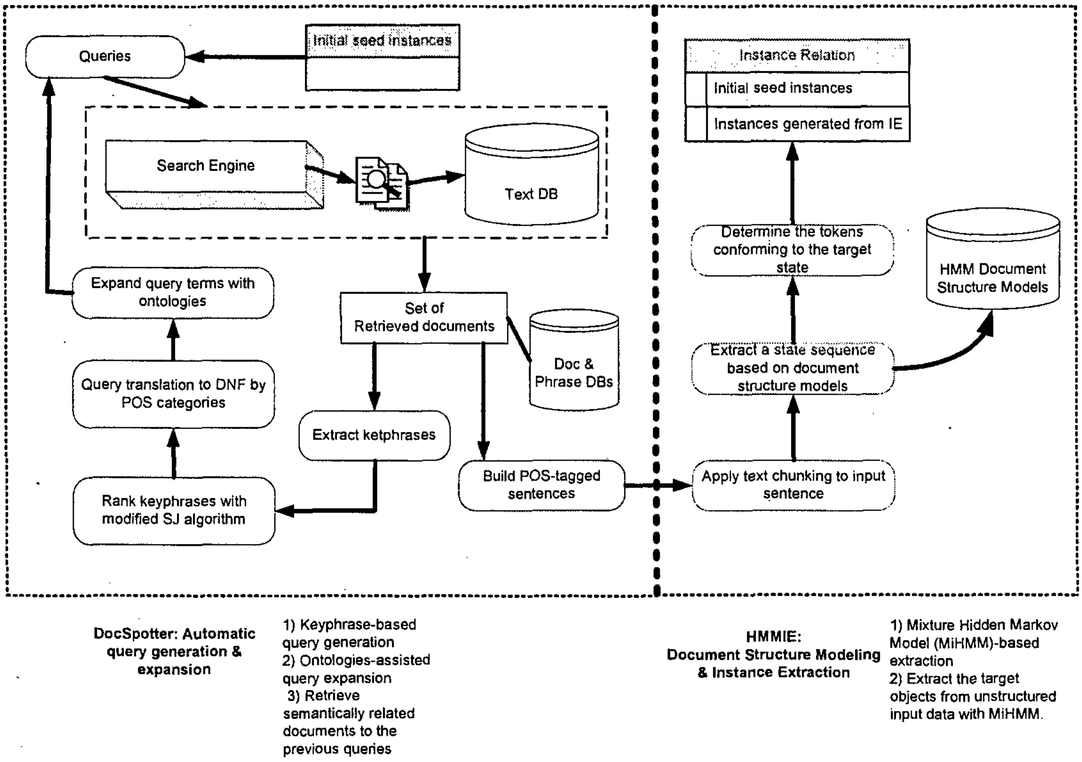


Figure 1: System architecture of RIKE

2. The instance set induces a binary partition (a split) on the documents: those that contain instances or those that do not contain any instance from the relation. The documents are thus labeled automatically as either positive or negative examples, respectively. The positive examples represent the documents that contain at least one instance. The negative examples represent documents that contain no instances.
3. RIKE next applies data mining and IR techniques to derive queries targeted to match—and retrieve— additional documents similar to the positive examples
4. RIKE then applies a HMM-based state sequence extraction technique over the documents. It models a set of document structures using the train documents. These models are kept in the model base, which will be used to serve as an engine for extracting state sequence from the documents.
5. The system queries the text databases using the automatically learned queries from Step 3 to retrieve a set of promising documents from the databases and then goes to Step 2. The whole procedure repeats until no new instances can be added into the relation or we reach the pre-set limit of a maximum number of text files to process.

The major contributions of this project are four-fold:

1. Automatic query generation for effective retrieval from biomedical text databases. This project introduces novel automatic query-based techniques to retrieve the articles/documents that are promising for the extraction of relations from text. The proposed technique is based on keyphrases and POS-tagged categorization. The keyphrases are extracted from the retrieved documents and weighted with an algorithm based on information gain and co-occurrence of phrases. It automatically discovers the characteristics of documents that are useful for extraction of a target relation and generates queries in each iteration to select potentially useful articles from the text databases.

2. A statistical generative model, Mixture Hidden Markov Model (MiHMM) for automatic relation extraction. This project proposes MiHMM, a mixture of Hidden Markov Models (HMMs), organized in a hierarchical structure to help the IE system cope with data sparseness. MiHMM takes a set of sentences with contextual cues that were identified by a Support Vector Machine-based text chunking technique. MiHMM then learns a generative probabilistic model of the underlying state transition structure of the sentence from a set of tagged training data. Given a trained probabilistic mixture model of the data, the system then applies this model to new unseen input documents to predict which portions of these documents are likely targets according to the training data template. MiHMM is different from existing HMM-based approaches as follows: (a) It employs probabilistic mixture of HMMs that is hierarchically structured. (b) It incorporates contextual and semantic cues into the learned models to extract knowledge from the unstructured text collections without any document structures. (c) It adopts a SVM text chunking technique to partition sentences into grammatically related groups. Thus using MiHMM for extracting biomedical entities has the following advantages over other approaches: (a) it overcomes the problem of the single POS HMMs with modeling the rich representation of text where features overlap among state units such as word, line, sentence, and paragraph. By incorporating sentence structures into the learned models, MiHMM provides better extraction accuracy than the single POS HMMs. (b) it resolves the issues with the single POS HMMs for IE that operate only on the semi-structured such as HTML documents and other text sources in which language grammar does not play a pivotal role.

3. Ontologies-assisted query expansion coupled with Keyphrases. In this project, we incorporates three ontologies such as WordNet, MESH, and UMLS into our query expansion technique. WordNet ontology is used to capture the noun form of the important verb terms such as "retrieve" or "interact" (Song et al., 2004). In our experiments, WordNet fails to make substantial enhancement of the retrieval performance. However, the experimental results showed that MeSH and UMLS made significant improvement of the retrieval performance. MeSH and UMLS ontologies are used to expand queries by providing synonymously associated phrases.

4. Minimum human intervention required for knowledge extraction. From a system architecture perspective, RIKE is designed and developed with design pattern and object-oriented design methodologies. The architectural soundness of RIKE supported by object-oriented paradigm makes it possible to implement minimum human-intervention operation of the system. The whole procedure proposed in this project was unsupervised with no human intervention except for a few seed instances provided by users in the very beginning. It makes it easy to port to a new domain without any changes to the extraction system. Also, it introduced a strategy for evaluating the quality of the patterns and the instances that are generated in the each iteration of the extraction process. Only those instances and patterns that are regarded as being "sufficiently reliable" were kept for the following iteration of the system. These new strategies for generating and filtering of patterns and instances improved the quality of the extracted instances and patterns significantly. In addition, RIKE is designed to plug into different search and extraction engines with minimum changes. RIKE is also ease to incorporate with various types of data formats such HTML, SGML, or XML.

Summary

We propose a novel knowledge extraction system, RIKE, which consists of DocSpotter, a keyphrase-based query expansion algorithm, and HiMMIE, a Mixture Hidden Markov model-based information extraction algorithm. We demonstrate that our query expansion algorithm based on keyphrases and the POS phrase category is effective and accurate in terms of average precision and precision at top 20. Encouraged by the previous studies on pseudo-relevance feedback, we apply keyphrase extraction techniques to query expansion. Along with keyphrase-based expansion, we employ a new word sense disambiguation technique in using ontology (e.g., WordNet) to add terms to the query. We also employ a POS phrase category-based Boolean constraint technique.

We also demonstrate that our techniques yield significant improvements over the well-established BM25 algorithm and SLIPPER for MEDLINE data on various search tasks including the protein-protein interaction tasks. Among four algorithms implemented, BM25, SLP, KP, and KP+C, the experimental results indicate that the KP+C algorithm proved to be the best.

We validate that ontologies-assisted query expansion improves the retrieval performance in the biomedical domain. We also prove that Mixture Hidden Markov model-based extraction is a high quality extraction algorithm, particularly useful for biomedical entity relations. We compare HiMMIE with four well-known IE techniques: 1) RAPIER, a rule-based machine learning system, 2) SVM, Support Vector Machine-based extraction algorithm, 3) MaxEnt, Maximum Entropy-based extraction algorithm, and 4) single POS HMM. Our experiments showed that HiMMIE outperforms these IE techniques in extracting protein-protein interactions in terms of F-measure.

P11. LSNMF: A modified non-negative matrix factorisation algorithm with embraced uncertainty measurements

Guoli Wang, Andrew V. Kossenkov, and Mike F. Ochs

Fox Chase Cancer Center

Background

Non-negative matrix factorisation (NMF), a machine learning algorithm, has been applied to the analysis of microarray data. A key feature of NMF is the ability to identify patterns that exist in a subset of the data. Microarray data generally includes individual estimates of uncertainty for each gene in each condition, however NMF does not use this information. Previous work has shown that such uncertainties can be extremely valuable for pattern recognition.

Results

We have extended NMF into a new algorithm called least squared non-negative matrix factorizatino, LS-NMF, which integrates uncertainty measurements of gene expression data into NMF updating rules. While the LS-NMF algorithm maintains advantages of original NMF algorithm, easy implementation and a guaranteed locally optimal solution, the performance in terms of grouping of genes into logical patterns has been significantly improved as demonstrated by ROC analysis. The area under the ROC curve for the Rosetta deletion mutant data set increased as much as 20% using LS-NMF in place of NMF. For predictive power of gene relationships, LS-NMF exceeds NMR by 12 – 40 times in terms of identifying functionally related genes as determined from the MIPS database.

Conclusion

Uncertainty measurements on gene expression data provide valuable information for data analysis, and integration of this information into the LS-NMF algorithm significantly improves the power of the NMF technique.

P12. A computational model for mammalian promoter identification

Junwen Wang and Sridhar Hannenhalli

Penn Center for Bioinformatics and Department of Genetics
University of Pennsylvania, Philadelphia, PA 19104-6021

Abstract

Background. An accurate identification of gene promoters remains an important challenge. Computational approaches for this problem rely on promoter sequence attributes that are believed to be critical for transcription initiation.

Methodology. Here we report a novel probabilistic model that captures two important properties of promoters, not used by previous methods, *viz.*, the location preference and co-occurrence of promoter elements. The derived promoter prediction method – *Position Specific Propensity Analysis (PSPA)* – is significantly more accurate than previously described methods. Additionally, many of the position-specific DNA elements are strongly linked with the function of the gene product. For instance, TATAA at 32 nucleotides upstream of the transcription start site is strongly linked with stress response network, cancer and apoptosis. A highly conserved novel motif CCTTT at -1 position is strongly associated with protein synthesis, cellular and tissue development. Our comparative analysis of promoter classes reveals that the promoters devoid of CpG islands are more conserved and have fewer alternative transcription start sites. The web server for the PSPA promoter predictor is available at http://cagr.pcbi.upenn.edu/PSPA.

Conclusions. Our work provides a better promoter model that can guide experiments to identify novel genes and novel transcription starts of known genes. The discovered links between promoter elements and gene function allows us to infer genetic networks from promoter elements. A greater conservation and fewer alternative transcription start sites in GC poor promoters may provide a mechanistic explanation for a tighter regulation of these genes as evidenced by their tissue restricted expression.

VI. DEMONSTRATIONS

DEMONSTRATION ABSTRACTS

D1. MollDE: a homology modeling framework you can click with.

A.A. Canutescu and R.L. Dunbrack Jr.
Fox Chase Cancer Center

SUMMARY: Molecular Integrated Development Environment (MolIDE) is an integrated application designed to provide homology modeling tools and protocols under a uniform, user-friendly graphical interface. Its main purpose is to combine the most frequent modeling steps in a semi-automatic, interactive way, guiding the user from the target protein sequence to the final three-dimensional protein structure. The typical basic homology modeling process is composed of building sequence profiles of the target sequence family, secondary structure prediction, sequence alignment with PDB structures, assisted alignment editing, side-chain prediction and loop building. All of these steps are available through a graphical user interface. MolIDE's user-friendly and streamlined interactive modeling protocol allows the user to focus on the important modeling questions, hiding from the user the raw data generation and conversion steps. MolIDE was designed from the ground up as an open-source, cross-platform, extensible framework. This allows developers to integrate additional third-party programs to MolIDE. AVAILABILITY:
http://dunbrack.fccc.edu/molide/molide.php

D2. PAINT: Promoter Analysis and Interaction Networks Toolset

Praveen Chakravarthula, Rajanikanth Vadegepalli, Gregory E. Gonye, and James S. Schwaber

Daniel Baugh Institute for Functional Genomics and Computational Biology, Department of Pathology, Thomas Jefferson University, Philadelphia.

Contact: Praveen Chakravarthula (praveenc@mail.dbi.tju.edu)

Abstract:

High throughput experimental technologies, such as microarrays, in conjunction with the availability of large biological databases, such as genome wide annotation, offer an ideal situation for the usage of computational methods and tools in the identification and analysis of gene regulatory networks. PAINT is an ongoing effort towards this end, that analyzes the potential promoters of a given gene list. Currently, PAINT supports the following mammalian systems: Human, Mouse and Rat, with future development plans to include other organisms.

Given a list of genes, PAINT can:

Fetch potential promoter sequences for the genes in the list.
Find Transcription Factor (TF) binding sites on the sequences.
Analyze the TF-binding site occurences for over/under-representation compared to a reference.
Output multiple visualizations for these analyses.
Generate hypotheses on significant Transcription Factors in the given context.

PAINT has been available as a web-based tool since 2003, at: http://www.dbi.tju.edu/dbi/tools/paint, with an update cycle of about 6 months, which includes database updates as well as program enchancements. The approach of PAINT has been applied in varied biological contexts, with interesting results, some of which have been experiementally validated.

Given the expanding scope of PAINT, we have been developing an interactive, integrative analysis framework, which we call Transcriptional Regulatory Network Analysis Workbench (TRNA Workbench) that incorporates the existing PAINT modules, with future plans to develop additional analysis components.

The TRNA Workbench is designed to organize an Analysis as a tree-like structure, that can accommodate analysis results from various Analysis Processors grouped under the appropriate categories. The Analysis Tree also facilitates a "Wizard" like step-by-step approach to applying an analysis methodology. Interaction with other programs is made possible by using the export functionality, which converts the results of each step available to text files.

The TRNA Workbench is currently beta software, and will be available for download on the PAINT web site (www.dbi.tju.edu/dbi/tools/paint/) in a couple of weeks. Written in Java, the Workbench currently runs on Windows, Linux and Mac OSX.

The poster will focus more on the TRNA Workbench (with PAINT modules), which will also be featured in the demonstration.

D3. Genomics Unified Schema (GUS) Web Development Kit (WDK): Building Data Mining Websites in Minutes

Steve Fischer, Thomas Gan, David Barkan*, Jerric Gao, Chris Stoeckert**

CBIL, PCBI, University of Pennsylvania, Philadelphia, PA 19104
QB3, University of San Francisco, San Francisco, CA 94143
CTEGD, University of Georgia, Athens, Georgia 30602

Genomics Unified Schema (GUS) is a functional genomics research platform originally developed at the Computational Biology and Informatics Laboratory (CBIL) at Penn. It includes a large relational schema and sophisticated application frameworks for data integration and data mining. As GUS became an open development project (http://gusdb.org) recently, a quickly growing number of bioinformatics labs and centers have adopted GUS. This has entailed a growing need to build data mining websites quickly.

We have developed GUS Web Development Kit (WDK, see http://gusdb.org/wdk/) to address such needs by the GUS community. WDK is centered around the concept of "Canned Queries", which are queries with parameters that can be posted on a website. It uses a Question-Summary-Record paradigm as the implementation. Questions are bindings of Queries (return rows) with Records (that are structured and strong typed data), while Summaries are list of summary attributes for Records returned by Questions. To build a website on top of a relational database (which does not have to be GUS), one only needs to specify in a model XML file the kind of Questions (with parameters) allowed on the website and the kinds of Records (with attribute populating queries) they return. WDK would take care of default site layout and site navigation control. Additional features supported by WDK include hooks for customizing any pages on the site (question page, summary page, record page etc.) or by Question/Record type and a query history with boolean expression for combining answers.

WDK comes with a set of command line tools for writing and testing the model XML file without going through the web server, and it also has a companion toySite (http://gusdb.org/wdk/toy/) that can be set up in a matter of minutes and used as a template for real sites.

D4. Evolution of Rural Biotech vs Refined Biotech

KBN Rayana., Director General-IAMMA, Center for Ag. Biotech., India /USA.
email: agrmktg@yahoo.com., kbnr2003@yahoocom

The Rural Biotech is the one of the foundation to advanced Biotech in the field of Biotechnology. The Biotech which makes the refind for further in the advancing of Biotech in recent days.

The start of Biotech from the Agriculture ie rural side will be discussed and advancement will be discussed in detail towards such advances in the food and Agricultural aspect.

The significant results which not only makes the advancements of the understanding the public but also do understandable for the advancements to the scientists with application process. The application is key role in the modern biotech which is hampering the application in the development process due to lack of application process and lack of advanced technology applied for field conditions etc.

Therefore the significance will be discussed.

VII. INDUSTRY RESEARCH PRESENTATIONS:

What Tough Bioinformatics/ Life Sciences Problems is Industry Tackling?

Session Chair: Guanghui Hu, Ph.D.

1:30 – 3:00

Bioinformatics in Translational Science: Bridging the Gap between pre-clinical and Clinical Discovery.

Anastasia M. Khoury Christianson, Ph. D.
Director, Discovery Medicine Informatics, AstraZeneca Pharmaceuticals

Abstract: Delivering better drugs to the market requires an understanding of disease mechanism during all stages of drug discovery, the availability of representative models of disease, both pre-clinical animal models and clinical / human models, the ability to relate pre-clinical data and results to clinical findings, and the availability of markers to measure disease stages, drug efficacy and safety, etc. In order to better understand the mechanism of disease and the drugs that act on them, we need to bridge the information gap between pre-clinical and clinical discovery. This presentation will summarize how Bioinformatics can contribute to the Translational Science/Translational Medicine area and will highlight the scientific, technical and cultural challenges of this work.

Target Identification through Expression Profiling and Pathway analysis, A Case Study

Yuchen Bai, Ph.D., Senior Research Scientist
Bioinformatics Client Services, Genomics,
Wyeth Research

Abstract

Recent advances in genomics technology have greatly expedited the drug discovery process in the pharmaceutical industry. The Wyeth Bioinformatics group within the Biotechnology department provides the infrastructure for biological data generation, storage, retrieval and analysis, and has made a tremendous impact on every aspect of the drug discovery research at Wyeth. Gene expression profiling and pathway analysis of tissues from normal and diseased individuals has proven to be a powerful approach not only to understand the intrinsic mechanisms of the disease, but also to identify genes that are potential drug targets for disease therapy and related applications.

In order to identify novel targets for female contraception, the Wyeth contraception biology group and the University of Pennsylvania initiated a collaboration to study the biology of unexplained infertility. Ovarian granulosa cells from women with mechanical (control) or unexplained infertility were collected during in vitro fertilization protocols and prepared for transcriptional profiling. Affymetrix human U133A arrays were used to generate gene expression data. An initial analysis found a dramatic variation of the scaling factor after global normalization and high expression of red blood cell markers in some infertile samples. The abundance of globin mRNA and associated ribosomal RNA in contaminating red blood cells greatly decreased the RNA complexity of the tissue samples. The log-transformed slope of the standard curve, which mimics the variation of the scale factor, was then used as a covariate in a linear statistical model to adjust for the effect of red blood cell contamination while assessing the effect of infertility. Differentially regulated genes in the infertile samples were studied in the context of gene networks and pathways by using the Ingenuity Pathway Analysis (IPA) tool. Many of the genes thus far identified and indicated in the networks have been known to playa role in normal ovarian function. The study will provide useful insights towards understanding the molecular mechanism underlying female infertility and identifying novel female contraception targets.

Integrative Biology: a mega-pixel view of disease and drug activity across species

Terence E Ryan PhD
Director of Integrative Biology
Discovery Research
GlaxoSmithKline
King of Prussia, PA

A challenge for post-genome bioinformatic analysis is the integration of other analyte classes into a systems biology viewpoint. By moving beyond nucleic acid and protein signatures, bioinformatics professionals have the opportunity to map the flux of lipids, metabolites, and other biomolecules across tissues and biofluids. Such a dynamic and systems-oriented approach will increase the breadth and depth of bioinformatic analyses. Resulting biomarkers that provide better insight into drug action in disease will help to fill a knowledge gap that contributes to translational efficacy failures resulting from unsuspected differences between animal models and human biology. GlaxoSmithKline's own experience with Integrative Biology and learnings from paired animal model and human experiments will be discussed.

VIII. PANEL DISCUSSION:

Academic/ Industry collaborations: Success Stories and Projects that Bombed – How? Why? Lessons Learned

Moderator: Wade Rogers, Ph. D.

1:30 – 3:00

Panel Moderator:

Wade Rogers, Ph.D., President & Chief Executive Officer, Cira Discovery Sciences, Inc.

Dr. Rogers is one of the three co-founders of Cira Discovery Sciences. Prior to founding Cira, Dr. Rogers was senior principal investigator at Bristol-Myers Squibb, where he led programs in computational approaches to target validation and lead compound discovery and optimization. He received a bachelor's of science in physics from the University of Delaware, a master's of science in engineering physics from the University of Virginia, and a doctorate in physics from the University of Colorado. He was a National Research Council Postdoctoral Fellow at the National Bureau of Standards (now NIST) before joining DuPont in 1981. While at DuPont he led the research team that developed forerunners to Cira's pattern discovery algorithms. He later left DuPont to join DuPont Pharmaceuticals.

Panelists:

Susan B. Davidson, Ph.D., Weiss Professor of Computer and Information Science, Deputy Dean of the School of Engineering and Applied Science, University of Pennsylvania

Dr. Davidson received the B.A. degree in Mathematics from Cornell University, Ithaca, NY, in 1978, and the M.A. and Ph.D. degrees in Electrical Engineering and Computer Science from Princeton University, Princeton NJ, in 1980 and 1982. Dr. Davidson is the Weiss Professor of Computer and Information Science and Deputy Dean of the School of Engineering and Applied Science at the University of Pennsylvania, where she has been since 1982.

Dr. Davidson's research interests include database and web-based systems, and bioinformatics. Within bioinformatics she is best known for her work with the Kleisli data integration system (joint work with Drs. Buneman, Tannen and Overton), which was subsequently commercialized in the company GeneticXChange. Her more recent work has centered on XML technologies for data sharing, data integration and data curation.

Dr. Davidson was the founding co-director of the Center for Bioinformatics (PCBI) from 1997-2000, and interim director from 2000-2003. The PCBI is a multi-school center which pulls together faculty from biomedicine, statistics, mathematics, engineering and computer science. The mission of the center is to foster research and education in the rapidly emerging fields of bioinformatics and computational biology, disciplines which deal with the management and analysis of data generated by high-throughput techniques in genomics, molecular and cellular biology.

She holds a secondary appointment in the Department of Genetics, is an (Association for Computing Machinery) ACM Fellow, received the Lenore Rowe Williams Award (2002), and was a Fulbright Scholar and recipient of a Hitachi Chair (2004).

Barbara Handelin, General Manager of Computational Biology & Director of Diagnostics, DNAPrint, Inc.

Dr. Handelin is a seasoned biotech business executive and board certified Medical Geneticist (Molecular and Biochemical Genetics) who has pioneered the responsible application of genetics to clinical medicine and drug development research over a 16 career. In 1987 Dr. Handelin established what became the largest commercial DNA testing laboratory in the world at Integrated Genetics (now Genzyme Genetics). After being recruited to found a gene therapy company in the Philadelphia region, Dr. Handelin founded her own consulting practice, Handelin Associates, in the biopharmaceutical and technology transfer sector. She delivered consulting services to venture capital investors, CEOs of new technology companies, senior business development executives of biotechnology, diagnostics, genomics, pharmacogenomics and bioinformatics companies and biomedical universities. In late 2000, Dr. Handelin founded a biosimulation company, Kenna Technologies, where she has served as CEO until the acquisition of Kenna by DNAPrint™ genomics Inc. She now serves as General Manager of Computational Biology and Director of Diagnostics at DNAPrint.

Dr. Handelin serves as Director at RedPath Integrated Pathology, and has served on technical and scientific advisory boards of several diagnostic and genomics companies. She also served for 10 years as a board member of the IRB educational nonprofit, Public Responsibility in Medicine and Research (PRIM&R). Dr. Handelin has served on a variety of federal committees and advisory panels on genetic testing and pharmacogenetics, including the Secretary's Advisory Panel on Genetics, Health and Society and was the Principal Investigator on a DOE ELSI grant on the "The Responsibility of Oversight in Genetics Research: How to Enable Effective Human Subjects Review of Public and Privately Funded Research Programs". Dr. Handelin took her Ph.D. at the Oregon Health Sciences University and the Massachusetts Institute of Technology and has authored journal publications in human genetics, bioethics for industry and genetics education.

Karl V. Steiner, Dr. Ing., Associate Director, Delaware Biotechnology Institute and Professor, Department of Electrical and Computer Engineering, University of Delaware

Dr. Steiner received his Engineering Doctorate from the University of Kaiserslautern in Germany, his Master's degree in Electrical and Computer Engineering from the University of Delaware and completed his undergraduate degree in Information Technologies in Braunschweig, Germany.

Dr. Steiner joined the University of Delaware in 1984. Prior to joining DBI in 2000, he served as the Executive Director of the University of Delaware Center for Composite Materials, an interdisciplinary research center in the College of Engineering, and one of the foremost academic research centers in its field. From 1996 to 1998 he was the Founding Executive Director of the Fraunhofer USA – Resource Center Delaware, a subsidiary of the German Fraunhofer Society, one of the largest non-profit applied research organizations in the world.

Dr. Steiner is currently the Program Coordinator of the NIH-NCRR funded IDeA Network of Biomedical Research Excellence (INBRE) and serves as Director of the Bioinformatics

Center under the INBRE program. He is also the Co-Principal Investigator for the NSF–EPSCoR Research Infrastructure Improvement (RII) Program with a focus on Complex Environmental Systems and Ecosystem Health. In his previous positions, he has led and coordinated numerous major multi-disciplinary programs, among them several Department of Defense supported Centers of Excellence.

Dr. Steiner has over twenty years of research experience in image enhancement and visualization methodologies. Much of his earlier research was focused on nondestructive evaluation and image analysis of engineered structures, such as aircraft wings, automotive panels, bridge structures, and hip implants. His current research interests are in the area of interactive immersive visualization methodologies for the life sciences, primarily in complex multi-variant data analysis and in biomedical imaging with a focus on virtual surgery simulations.

Dr. Steiner has contributed over 50 technical publications in international journals and conference proceedings related to manufacturing science, nondestructive evaluation, computer visualization, and image analysis methodologies. He has served as conference and session chair in numerous technical conferences and workshops and as reviewer for several journals focused on composite materials and nondestructive testing. He has given many research seminars and presentations at international conferences and countless talks and lectures at industrial sites and academic institutions across the world.

Jeffrey S. Wiseman, Ph.D., Vice President, Technology and Informatics, Locus Pharmaceuticals, Inc.

Dr. Wiseman joined Locus in June 2003 with 24 years of experience in the pharmaceutical industry. Dr. Wiseman was most recently Global Vice President of Cheminformatics at GlaxoSmithKline (GSK). While at GSK, Dr. Wiseman's division provided the informatics underlying the "industrialization" of GSK's worldwide Discovery Research high-throughput activities, including the design of GSK's 1.3 million chemical compound library collection, ultra high-throughput screening and high-throughput chemistry. Prior to this role, Dr. Wiseman held key management positions in research computing and molecular sciences at both Glaxo Wellcome and SmithKline Beecham. Dr. Wiseman started his industry career at the Merrell Dow Research Institute where he was Group Leader of Enzyme Chemistry.

Dr. Wiseman holds a Ph.D. in Chemistry from Harvard University, a B.S. in Chemistry from Ohio University, and trained as a Postdoctoral Research Fellow at Stanford University and Brandeis University prior to commencing his career in the pharmaceutical industry.

IX. LIST OF PARTICIPANTS

Name	Affiliation	Email
Sankar Addya	Thomas Jefferson University Hospital	sankar.addya@jefferson.edu
David Austin	Austin	daustin@mail.med.upenn.edu
Amit Bahl	University of Pennsylvania	abahl@mail.med.upenn.edu
Yuchen Bai	Wyeth Pharmaceuticals	baiy@wyeth.com
Bubu Banini	Thomas Jefferson University	Bubu.Banini@jefferson.edu
Jeff Barton	N/A	bartojn@comcast.net
Ghislain Bidaut	University of Pennsylvania	ghbidaut@pcbi.upenn.edu
Alex Birch	Merck	abirch@gmail.com
Dan Blankenberg	Penn State	djb396@psu.edu
Irina Bochkis	University of Pennsylvania	ibochkis@mail.med.upenn.edu
James Brown	GlaxoSmithKline	james.r.brown@gsk.com
Brian Brunk	University of Pennsylvania	brunkb@pcbi.upenn.edu
Zivjena Buletic	Temple University Medical School	zbuletic@temple.edu
Adrian Canutescu	Fox Chase Cancer Center	Adrian.Canutescu@fccc.edu
Bryan Cardillo	University of Pennsylvania	dillo@seas.upenn.edu
Praveen Chakravarthula	Daniel Baugh Institute, Thomas Jefferson	praveenc@mail.dbi.tju.edu
Hareesh Chandrupatla	University of the Sciences	hchandru@yahoo.com
Gary Guang Chen	CBIL, PCBI, UPENN	ggchen@pcbi.upenn.edu
Feng Chen	University of Pennsylvania	fengchen@sas.upenn.edu
Anastasia Christianson	AstraZeneca	anastasia.christianson@astrazeneca.com
Wen-Yu Chung	PSU	wuc114@psu.edu
Shirley Cohen	University of Pennsylvania	shirleyc@seas.upenn.edu
John Condon	PSGV	jmhcondon@hotmail.com
Shailesh Date	University of Pennsylvania	svdate@pcbi.upenn.edu
Susan Davidson	University of Pennsylvania	susan@cis.upenn.edu
Margaret DiFilippo	Ingenuity	mdifilippo@ingenuity.com
Sameepa Doshi	University of the Sciences in Philadelphia	sameepadoshi@gmail.com
Roland Dunbrack	Fox Chase Cancer Center	Roland.Dunbrack@fccc.edu
Ekene	USIP	eavery@usip.edu
Tolga Emre	Wistar Institute	emre@sas.upenn.edu
Kobby Essien	University of Pennsylvania	kobby@seas.upenn.edu
Jan Feng	Temple University	feng@temple.edu
Dirk Fey	DBI, Thomas Jefferson University	dirkfey@,ail.dbi.tju.com
Richelle Francis	USDA / USiP	RFrancis@errc.ars.usda.gov
Xiaowu Gai	The Children's Hospital of Philadelphia	bioinformatician@hotmail.com
Stephen Gallagher	UPENN	sgallagh@cceb.med.upenn.edu
Thomas Gan	UPenn	ygan@pcbi.upenn.edu
Ge	CHOP	zhangg@email.chop.edu
Hugo Geerts	In Silico Biosciences	Hugo-Geerts@In-Silico-Biosciences.com
Gregory Gonye	Thomas Jefferson University	ggonye@mail.dbi.tju.edu
Michael Gormley	Drexel University	mpg33@drexel.edu
Regina Gorski	University of Pennsylvania	rgorski@mail.med.upenn.edu
Jay Greenberg	USIP	greenbergj@hotmail.com

Gerald B. Grunwald, Ph.D.	Thomas Jefferson University	gerald.grunwald@jefferson.edu
Dexter Hadley	University of Pennsylvania	dexter@mail.med.upenn.edu
Adam Halasz	University of Pennsylvania	halasz@grasp.upenn.edu
John Hall	LifeSensors, Inc.	hall@lifesensors.com
Bo Han	IST, Temple University	hanbo@ist.temple.edu
Haiping Hao	Thomas Jefferson University	hhao@mail.dbi.tju.edu
Ned Haubein	The BioAnalytics Group	ned@bioanalyticsgroup.com
Tammy Heesakker	GPBA	theesakker@bioadvance.com
Amanda Herz	Penn State Great Valley	ach14@psu.edu
John G. Hoey	Ingenuity Systems	jhoey@ingenuity.com
Steve Hookway	Lehigh University	snh3@lehigh.edu
Guanghui Hu	GlaxoSmithKline	guanghui.2.hu@gsk.com
Xiaohua Tony Hu	Drexel University	thu@cis.drexel.edu
Chris Huang	Centocor R&D Inc.	chuang4@cntus.jnj.com
Heshu Huang	Thomas Jefferson University	heshu.huang@jefferson.edu
Marcin Imielinski	University of Pennsylvania School	imielins@mail.med.upenn.edu
John Iodice	Penn Center for Bioinformatics	iodice@pcbi.upenn.edu
Yang Jin	University of Pennsylvania	yajin@mail.med.upenn.edu
Mary Ann Jones	New Jersey Institute of Technology	majones@coopknow.com
Tom Kappeler	DSI	tkappeler@hotmail.com
Rishi Khan	Thomas Jefferson University	rishi@mail.dbi.tju.edu
Val Kogan	Mid-Atlantic - Russia Business Council	val@ma-rbc.org
Edward Kogan	Mid-Atlantic - Russia Business Council	pfalconed@yahoo.com
Despina Kontos	Temple University - CIS Dept.	dkontos@temple.edu
Andrei Kossenkov	Fox Chase Cancer Center & Drexel	av_kossenkov@fccc.edu
Shubhada Kulkarni	Ms	shubhada.kulkarni@firstdata.com
Yvonne Le	PHA	mai.lan9999@gmail.com
Jun Li	GlaxoSmithKline	jun_u_li@gsk.com
Wei Li	Temple University	weiwei61@temple.edu
Yanhong Liu	USDA	yliu@errc.ars.usda.gov
Jaron Liu	Lifesensors Inc.	jaronster@gmail.com
Daniel Lopresti	Lehigh University	lopresti@cse.lehigh.edu
Aaron Mackey	Univ. of Pennsylvania	amackey@pcbi.upenn.edu
Michal Magid-Slav	GlaxoSmithKline	Michal_M_Magid-Slav@gsk.com
Elisabetta Manduchi	University of Pennsylvania	manduchi@pcbi.upenn.edu
Joan Mazzarelli	University of Pennsylvania	mazz@pcbi.upenn.edu
Suzanne McCahan	Thomas Jefferson University	smmccaha@nemours.org
Uros Midic	Temple University	uros@ist.temple.edu
David Mosenkis	GPBA Consultant	dmosenkis@gmail.com
Allan Moser	Cira Discovery Sciences	allan.moser@ciradiscovery.com
Michael J. Moser	Penn State Great Valley	Piovoso@psu.edu
Philip Mui	GlaxoSmithKline	philip_w_mui@gsk.com
Muhammad Mukhtar	Thomas Jefferson University	muhammad.mukhtar@jefferson.edu
Dennis Murphy	GlaxoSmithKline	dennis.6.murphy@gsk.com
Sarita Nair	University of the Sciences in Philadelphia	sarita@mail.dbi.tju.edu
Syam Ameya Namballa	University of Sciences in Philadelphia	syamameya@yahoo.com

Jeffrey C. Nash	DeVry University	jnash@phi.devry.edu
Mohammed Nayeem	None	asker@excite.com
Charles Nicholson	Chadless Enterprises	gpba@chadless.net
NoDocs	NoDocs	NoDocs@test.com
Zoran Obradovic	Temple University	zoran@ist.temple.edu
Michael Ochs	Fox Chase Cancer Center	m_ochs@fccc.edu
Zahida Parveen	Thomas Jefferson University	zahida.parveen@jefferson.edu
Amit Patel	University of the Sciences of Philadelphia	amitupatel@gmail.com
Saurav Pathak	University of Pennsylvania	saurav@grasp.upenn.edu
Kimberly Pearson	CHOP	kpearson17@comcast.net
Lucia Peixoto	UPENN	luciap@sas.upenn.edu
Juan Carlos Perin	Children's Hospital of Phil.	bic@genome.chop.edu
Tom Petty	University of Pennsylvania - GCB	tpetty@mail.med.upenn.edu
Pina	USDA	pfratamico@errc.ars.usda.gov
Deborah Pinney	Univ Penn	pinney@pcbi.upenn.edu
Piyush	Drexel University	piyush.arora@drexel.edu
Angel Pizarro	ITMAT, Univ of PA	angel@mail.med.upenn.edu
Craig Pratt	Thomas Jefferson Unviversity	Craig.Pratt@jefferson.edu
Ramez	University of Pennsylvania	rzakhary@ldc.upenn.edu
Eric Rappaport	Children's Hospital of Philadelphia	rappaport@email.chop.edu
KBN Rayana	IAMMA	kbnr@hotmail.com
Steven Rayana	C.H.O.P. / Rutgers	carroll@genome.chop.edu
Brendan Reilly	TJU/Penn	ab3152524632@sec.com
Isidore Rigoutsos	IBM TJ Watson Research Center	rigoutso@us.ibm.com
David Russell	Penn State Great Valley	drussell@psu.edu
Chris Sarnowski	University of Pennsylvania	csarnows@pcbi.upenn.edu
Genaro Scavello	GenoVision, Inc	genaro.scavello@genovision.com
Walter Scheirer	Lehigh University	wjs3@lehigh.edu
Ms. Shoba Sharma	Software Consultant	shoba_sharma@yahoo.com
Ariel Shatz	FreeMind	ariel@freemind.co.il
Daniel F. Simola	University of Pennsylvania	simola@mail.med.upenn.edu
Randall F Smith	GlaxoSmithKline Bioinformatics	Randall_F_Smith@gsk.com
Jeffrey K. Smith	Berkery, Noyes & Co.	jeff.smith@berkerynoyes.com
Shep Smithline	Linuxetworx	shep@usinternet.com
Min Song	Temple University	min.song@temple.edu
Chris Stoeckert	Penn Center for Bioinformatics	stoeckrt@pcbi.upenn.edu
Ted Stolarczyk	SAS Institute Inc.	ted.stolarczyk@sas.com
Mark Strohmaier	Lehigh University, Computer Science Dept	mws205@lehigh.edu
Robert Styer	Villanova University	robert.styer@villanova.edu
Saul Surrey	Jefferson	saul.surrey@jefferson.edu
Thomas	Drexel	tcloy@yahoo.com
John Tobias	Univesity of Pennsylvania	jtobias@pcbi.upenn.edu
Oleh Tretiak	Drexel University	tretiak@coe.drexel.edu
John Ulicny	Temple Hospital	john.ulicny@tuhs.temple.edu
Raj Vadigepalli	Thomas Jefferson University	raj@mail.dbi.tju.edu
Alexandra Vamvakidou	Drexel University	av54@drexel.edu

Subbiah Venkatachala	Siemens	subbiah.venkatachalam@siemens.com
Pam Vercellone-Smith	Penn State Great Valley	pav115@psu.edu
Slobodan Vucetic	Temple University	vucetic@ist.temple.edu
Samir Wadhawan	Penn State University	srw194@psu.edu
Vladimir Walko	Medical Growth Partners	medgrow@msn.com
Junwen Wang	PCBI, UPENN	junwen@pcbi.upenn.edu
Guoli Wang	Fox Chase Cancer Center	GL_Wang@fccc.edu
Pei Wang	Temple University	pei.wang@temple.edu
Winfield Watson	LifeCycle Software	Winfield.Watson2@verizon.net
Laurence Weinberger	Lipton, Weinberger & Husick	larry@lawpatent.com
Adam Wenocur	University of the Sciences in Philadelphia	awenocur@usip.edu
Pete White	CHOP	white@genome.chop.edu
Peter White	University of Pennsylvania	pwhite@mail.med.upenn.edu
Christopher Wojciechowski	Lehigh University	cjw8@lehigh.edu
Weichen Wu	University of Pensylvania	wewu@seas.upenn.edu
Hsiuan-Lin Wu	Rutgers University	hsiuan@eden.rutger.edu
Daniel Wu	Drexel University	daniel.wu@drexel.edu
Michael Xie	Temple University	hongbox@temple.edu
Qing Xie	GlaxoSmithKline	qing.2.xie@gsk.com
Guochun Xie	Merck & Co., inc.	guochun_xie@merck.com
Ramez Zakhary	University of Pennsylvania	rzakhary@ldc.upenn.edu
Dalal Zakhary	University of Pennsylvania	zakhary@ldc.upenn.edu
Zhe Zhang	University of Pennsylvania	zhangz@email.unc.edu
Zhe Zhang	University of Pennsylvania	zhezhang@mail.med.upenn.edu
Yan Zhou	Fox Chase Cancer Center	yan.zhou@fccc.edu
Xun Zuo	LifeSensors Inc.	zuo@lifesensors.com